Volatility Forecast Comparison using Imperfect Volatility Proxies

Andrew J. Patton*
London School of Economics

First version: March 2004. This version: 29 April, 2006.

Abstract

The use of a conditionally unbiased, but imperfect, volatility proxy can lead to undesirable outcomes in standard methods for comparing conditional variance forecasts. We derive necessary and sufficient conditions on functional form of the loss function for the ranking of competing volatility forecasts to be robust to the presence of noise in the volatility proxy, and derive some interesting special cases of this class of "robust" loss functions. We motivate the theory with analytical results on the distortions caused by some widely-used loss functions, when used with standard volatility proxies such as squared returns, the intra-daily range or realised volatility. The methods are illustrated with an application to the volatility of returns on IBM over the period 1993 to 2003.

Keywords: forecast evaluation, forecast comparison, loss functions, realised variance, range.

J.E.L. Codes: C53, C52, C22.

*The author would particularly like to thank Peter Hansen, Ivana Komunjer and Asger Lunde for helpful suggestions and comments. Thanks also go to Torben Andersen, Tim Bollerslev, Rob Engle, Christian Gourieroux, Tony Hall, Mike McCracken, Nour Meddahi, Roel Oomen, Adrian Pagan, Neil Shephard, Kevin Sheppard, and Ken Wallis. Runquan Chen provided excellent research assistance. The author gratefully acknowledges financial support from the Leverhulme Trust under grant F/0004/AF. Some of the work on this paper was conducted while the author was a visiting scholar at the School of Finance and Economics, University of Technology, Sydney. Contact address: Financial Markets Group, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom. Email: a.patton@lse.ac.uk. Matlab code used in this paper is available from http://fmg.lse.ac.uk/∼patton/research.html.

1 Introduction

Many forecasting problems in economics and finance involve a variable of interest that is unobservable, even *ex post*. The most prominent example of such a problem is the forecasting of volatility for use in financial decision-making. Other problems include forecasting the true rates of inflation, GDP growth or unemployment (not simply the announced rates); forecasting trade intensities; and forecasting default probabilities or 'crash' probabilities. While evaluating and comparing economic forecasts is a well-studied problem, dating back at least to Cowles (1933), if the variable of interest is latent then the problem of forecast evaluation and comparison becomes more complicated[1].

This complication can be resolved, at least partly, if a conditionally unbiased estimator of the latent variable of interest is available. In volatility forecasting, for example, the squared return on an asset over the period t (assuming a zero mean return) is a conditionally unbiased estimator of the true unobserved conditional variance of the asset over the period t.[2] Many of the standard methods for forecast evaluation and comparison, such as the Mincer-Zarnowitz (1969) regression and the Diebold and Mariano (1995) and West (1996) tests, can be shown to be applicable when such a conditionally unbiased proxy is used, see Hansen and Lunde (2006) for example. However, it is not true that using a conditionally unbiased proxy will *always* lead to the same outcome as if the true latent variable was used, as shown Andersen and Bollerslev (1998), Andersen, *et al.* (2005a) and Hansen and Lunde (2006). In particular, some of the methods employed in recent applied work can lead to perverse outcomes.

For example, in the volatility forecasting literature numerous authors have expressed concern that a few extreme observations may have an unduly large impact on the outcomes of forecast evaluation and comparison tests, see Bollerslev and Ghysels (1994), Andersen, *et al.* (1999) and Poon and Granger (2003) amongst others. One common response to this concern is to employ forecast loss functions that are "less sensitive" to large observations than the usual squared forecast error loss function, such as absolute error or proportional error loss functions. In this paper we show analytically that such approaches can lead to incorrect inferences and the selection of inferior forecasts over better forecasts.

We focus on volatility forecasting as a specific case of the more general problem of latent variable forecasting. In Section 5 we discuss the extension of our results to other latent variable forecasting

[1] For recent surveys of the forecast evaluation literature see Clements (2005) and West (2005). For recent surveys of the volatility forecasting literature, see Andersen, *et al.* (2005b), Poon and Granger (2003) and Shephard (2005).

[2] The high/low range and realised volatility, see Parkinson (1980) and Andersen, *et al.* (2003) for example, have also been used as volatility proxies.

problems. Our research builds on work by Andersen and Bollerslev (1998), Meddahi (2001) and Hansen and Lunde (2006), who were among the first to analyse the problems introduced by the presence of noise in a volatility proxy. This paper is most closely related to the paper of Hansen and Lunde (2006), and we extend their work in two important directions: Firstly, we derive explicit analytical results for the undesirable outcomes that may arise when some common loss functions are employed, considering the three most commonly-used volatility proxies: the daily squared return, the intra-daily range and a realised variance estimator, and show that the distortions vary greatly with the choice of loss function. Secondly, we provide necessary and sufficient conditions on the functional form of the loss function to ensure that the ranking of various forecasts is preserved when using a noisy volatility proxy. These conditions are related to those of Gourieroux, *et al.* (1984) for quasi-maximum likelihood estimation[3].

The canonical problem in point forecasting is to find the forecast that minimises the expected loss, conditional on time t information. That is,

$$\hat{Y}^*_{t+h,t} \equiv \arg\min_{\hat{y}\in\mathcal{Y}} E\left[L\left(Y_{t+h}, \hat{y}\right)|\mathcal{F}_t\right] \tag{1}$$

where Y_{t+h} is the variable of interest, L is the forecast user's loss function, $\mathcal{Y}$ is the set of possible forecasts, and $\mathcal{F}_t$ is the time t information set. Starting with the assumption that the forecast user is interested in the conditional variance, and that some noisy volatility proxy will be used in evaluation tests, we effectively take the *solution* of the optimisation problem above (the conditional variance) as given, and consider the loss functions that will generate the desired solution. This approach is unusual in the economic forecasting literature: the more common approach is to take the forecast user's loss function as given and derive the optimal forecast for that loss function; related papers here are Granger (1969), Engle (1993), Christoffersen and Diebold (1997), Christoffersen and Jacobs (2004) and Patton and Timmermann (2004), amongst others. The fact that we know the forecast user desires a variance forecast places limits on the class of loss functions that may be used for volatility comparison, ruling out some choices previously used in the literature. However we show that the class of "robust" loss functions still admits a wide variety of loss functions, allowing much flexibility in representing volatility forecast users' preferences.

[3] All of the results in this paper apply directly to the problem of forecasting integrated variance, which Andersen, *et al.* (2002), amongst others, argue is a more "relevant" notion of variability. In that application, we take expected integrated variance rather than the conditional variance as the latent object of interest, and we require that an unbiased realised variance estimator is available. We focus on the problem of conditional variance forecasting due to its prevalence in applied work in the past two decades.

One of the main practical findings of this paper is that the stated goal of forecasting the conditional variance is not consistent with the use of some loss functions when an imperfect volatility proxy is employed. However, these loss functions are not themselves inherently invalid or inappropriate: if the forecast user's preferences are indeed described by an "non-robust" loss function, then this simply implies that the object of interest to that forecast user is not the conditional variance but rather some other quantity[4]. If the object of interest to the forecast user is known to be the conditional variance then this paper outlines tests for forecast comparison that are applicable when an imperfect volatility proxy is employed.

The remainder of this paper is as follows. In Section 2 we analytically consider volatility forecast comparison tests using an imperfect volatility proxy, showing the problems that arise when using some common loss functions. We initially consider using squared daily returns as the proxy, and then consider using the range and realised variance. In Section 3 we provide necessary and sufficient conditions on the functional form of a loss function for the ranking of competing volatility forecasts to be robust to the presence of noise in the volatility proxy, and derive some interesting special cases of this class of robust loss functions. One of these special cases is a parametric family of loss functions that nests two of the most widely-used loss functions in the literature, namely the MSE and QLIKE loss functions. In Section 4 we present an illustration using two widely-used volatility models, and in Section 5 we conclude and suggest extensions. All proofs and derivations are provided in appendices.

1.1 Notation

Let r_t be the variable whose conditional variance is of interest, usually a daily or monthly asset return in the volatility forecasting literature. Let the information set used in the forecasts be denoted $\mathcal{F}_{t-1}$, which is assumed to contain $\sigma\left(r_{t-j}, j \geq 1\right)$, but may also include other variables and/or variables measured at a higher frequency than r_t (such as intra-daily returns). Denote $V\left[r_t|\mathcal{F}_{t-1}\right] \equiv V_{t-1}\left[r_t\right] \equiv \sigma_t^2$. We will assume throughout that $E\left[r_t|\mathcal{F}_{t-1}\right] \equiv E_{t-1}\left[r_t\right] = 0$, and so $\sigma_t^2 = E_{t-1}\left[r_t^2\right]$. Let $\varepsilon_t \equiv r_t/\sigma_t$ denote the 'standardised return'. Let a forecast of the conditional variance of r_t be denoted h_t, or $h_{i,t}$ if there is more than one forecast under analysis. We will take forecasts as "primitive", and not consider the specific models and estimators that may have

[4] For example, the utility of realised returns on a portfolio formed using a volatility forecast, or the profits obtained from an option trading strategy based on a volatility forecast, see West, *et al.* (1993) and Engle, *et al.* (1993) for example, define economically meaningful loss functions, even though the optimal forecasts under those loss functions will not generally be the true conditional variance.

generated the forecasts. The loss function of the forecast user is $L : \mathbb{R}_+ \times \mathcal{H} \to \mathbb{R}_+$, where the first argument of L is σ_t^2 or some proxy for σ_t^2, denoted $\hat{\sigma}_t^2$, and the second is h_t. $\mathbb{R}_+$ and $\mathbb{R}_{++}$ denote the non-negative and positive parts of the real line respectively, and $\mathcal{H}$ is a compact subset of $\mathbb{R}_{++}$. Commonly used volatility proxies are the squared return, r_t^2, realised volatility, RV_t, and the range, RG_t. Optimal forecasts for a given loss function will be denoted h_t^* and are defined as:

$$h_t^* \equiv \underset{h \in \mathcal{H}}{\arg\min}\ E\left[L\left(\hat{\sigma}_t^2, h\right) | \mathcal{F}_{t-1}\right] \tag{2}$$

2 Volatility forecast comparison using an imperfect volatility proxy

We consider volatility forecast comparison tests based on (unconditional) expected loss, based on the work of Diebold and Mariano (1995) and West (1996). If we define $u_{i,t} \equiv L\left(\sigma_t^2, h_{i,t}\right)$, where L is the forecast user's loss function, and let $d_t = u_{1,t} - u_{2,t}$, then a DMW test of equal predictive accuracy can be conducted as a simple Wald test that $E[d_t] = 0$.[5]

Of primary interest is whether the feasible ranking of two forecasts obtained using an imperfect volatility proxy is the same as the infeasible ranking that would be obtained using the unobservable true conditional variance. We define loss functions that yield such an equivalence as "robust":

Definition 1 *A loss function, L, is "robust" if the ranking of any two (possibly imperfect) volatility forecasts, h_{1t} and h_{2t}, by expected loss is the same whether the ranking is done using the true conditional variance, σ_t^2, or some conditionally unbiased volatility proxy, $\hat{\sigma}_t^2$. That is,*

$$E\left[L\left(\sigma_t^2, h_{1t}\right)\right] \gtreqless E\left[L\left(\sigma_t^2, h_{2t}\right)\right] \Leftrightarrow E\left[L\left(\hat{\sigma}_t^2, h_{1t}\right)\right] \gtreqless E\left[L\left(\hat{\sigma}_t^2, h_{2t}\right)\right] \tag{3}$$

Meddahi (2001) showed that the ranking of forecasts on the basis of the R^2 from the Mincer-Zarnowitz regression:

$$\hat{\sigma}_t^2 = \beta_0 + \beta_1 h_{it} + e_{it} \tag{4}$$

is robust to noise in $\hat{\sigma}_t^2$. Hansen and Lunde (2006) showed that the R^2 from a regression of $\log\left(\hat{\sigma}_t^2\right)$ on a constant and $\log\left(h_t\right)$ is not robust to noise, and showed more generally that a sufficient condition for a loss function to be robust is that $\partial^2 L\left(\sigma^2, h\right) / \partial\left(\sigma^2\right)^2$ does not depend on h_t. In Section 3

[5] The key difference between the approaches of Diebold and Mariano (1995) and West (1996) is that the latter explicitly allows for forecasts that are based on estimated parameters, whereas the null of equal predictive accuracy is based on population parameters, see West (2005). The problems we identify below arise even in the absence of estimation error in the forecasts, thus our treatment of the forecasts as primitive, and so for our purposes these two approaches coincide.

we generalise this result by providing necessary and sufficient conditions for a loss function to be robust.[6,7]

It is worth noting that although the ranking obtained from a robust loss function will be invariant to noise in the proxy, the actual level of expected loss obtained using a proxy will be larger than that which would be obtained when using the true conditional variance. This point was compellingly presented in Andersen and Bollerslev (1998) and Andersen, *et al.* (2004). Andersen, *et al.* (2005a) provide a method to estimate the distortion in the level of expected loss and thereby obtain an estimator of the level of expected loss that would be obtained using the true latent variable of interest.

Notice that for any robust loss function the true conditional variance is the optimal forecast (we formally show this in the proof of Proposition 2), and thus a necessary condition for a loss function to be robust to noise is that the true conditional variance is the optimal forecast. In this section we determine whether this condition holds for some common loss functions, and analytically characterise the distortion for those cases where it is violated.

Under squared-error loss, also known as MSE loss, one can easily show that the optimal forecast is the conditional variance: $h_t^* = E_{t-1}\left[\hat{\sigma}_t^2\right] = \sigma_t^2$. Thus a DMW comparison of the true conditional variance with any other volatility forecast, using a conditionally unbiased volatility proxy and MSE as the loss function, will lead to the selection of the true conditional variance, subject to sampling variability. Further, it is clear that the MSE loss function also satisfies the sufficient condition of Hansen and Lunde (2006), and thus MSE is a "robust" loss function.

One common response to the concern that a few extreme observations drive the results of volatility forecast comparison studies is to employ alternative measures of forecast accuracy, see Pagan and Schwert (1990), Bollerslev and Ghysels (1994), Bollerslev, *et al.* (1994), Diebold and Lopez (1996), Andersen, *et al.* (1999), Poon and Granger (2003) and Hansen and Lunde (2005), for example. A collection of loss functions employed in the literature on volatility forecast evaluation and comparison is presented below. Some of these loss functions are called different names by different authors: MSE-prop is also known as "heteroskedasticity-adjusted MSE (HMSE)"; MAE-

[6] Our use of "robust" is related, though not equivalent, to the use of this adjective in estimation theory, where it applies to estimators that insensitive/less sensitive to the presence of outliers in the data, see Huber (1981) for example. A "robust" loss function, in the sense of Definition 1, will generally not be robust to the presence of outliers.

[7] We focus on measures of accuracy that can be expressed as sample means of losses incurred on each period in the sample. Rankings based on R^2 from regressions do not fit within this framework. See Hansen and Lunde (2006) for more discussion of R^2 as a ranking criterion.

prop is also known as "mean absolute percentage error (MAPE)" or as "heteroskedasticity-adjusted MAE (HMAE)".

$$
\begin{aligned}
MSE &: L\left(\hat{\sigma}_t^2, h_t\right) = \left(\hat{\sigma}_t^2 - h_t\right)^2 && (5)\\
QLIKE &: L\left(\hat{\sigma}_t^2, h_t\right) = \log h_t + \frac{\hat{\sigma}_t^2}{h_t} && (6)\\
MSE\text{-}LOG &: L\left(\hat{\sigma}_t^2, h_t\right) = \left(\log \hat{\sigma}_t^2 - \log h_t\right)^2 && (7)\\
MSE\text{-}SD &: L\left(\hat{\sigma}_t^2, h_t\right) = \left(\hat{\sigma}_t - \sqrt{h_t}\right)^2 && (8)\\
MSE\text{-}prop &: L\left(\hat{\sigma}_t^2, h_t\right) = \left(\frac{\hat{\sigma}_t^2}{h_t} - 1\right)^2 && (9)\\
MAE &: L\left(\hat{\sigma}_t^2, h_t\right) = \left|\hat{\sigma}_t^2 - h_t\right| && (10)\\
MAE\text{-}LOG &: L\left(\hat{\sigma}_t^2, h_t\right) = \left|\log \hat{\sigma}_t^2 - \log h_t\right| && (11)\\
MAE\text{-}SD &: L\left(\hat{\sigma}_t^2, h_t\right) = \left|\hat{\sigma}_t - \sqrt{h_t}\right| && (12)\\
MAE\text{-}prop &: L\left(\hat{\sigma}_t^2, h_t\right) = \left|\frac{\hat{\sigma}_t^2}{h_t} - 1\right| && (13)
\end{aligned}
$$

2.1 Using squared returns as a volatility proxy

In this section we will focus on the use of daily squared returns for volatility forecast evaluation, and in Section 2.2 we will examine the use of realised volatility and the range. We will derive our results under three assumptions for the conditional distribution of daily returns:

$$
r_t|\mathcal{F}_{t-1} \sim \begin{cases} F_t\left(0, \sigma_t^2\right) \\ \textit{Student's } t\left(0, \sigma_t^2, \nu\right) \\ N\left(0, \sigma_t^2\right) \end{cases}
$$

where $F_t\left(0, \sigma_t^2\right)$ is some unspecified distribution with mean zero and variance σ_t^2,and *Student's* $t\left(0, \sigma_t^2, \nu\right)$ is a Student's t distribution with mean zero, variance σ_t^2 and ν degrees of freedom. In all cases it is clear that $E_{t-1}\left[r_t^2\right] = \sigma_t^2$, and so the squared daily return is a valid volatility proxy.

Above we showed that the MSE loss function satisfied the necessary condition, that the optimal forecast is the true conditional variance. Now consider the MAE loss function from above. As usual with an absolute-error loss function we obtain the median as the optimal forecast:

$$
\begin{aligned}
h_t^* &= Median_{t-1}\left[r_t^2\right] \\
&= \sigma_t^2 \cdot Median_{t-1}\left[\varepsilon_t^2\right] \\
&= \begin{cases} \sigma_t^2 \cdot \frac{\nu-2}{\nu} \cdot Median\left[F_{1,\nu}\right], & \text{if } r_t|\mathcal{F}_{t-1} \sim \textit{Student's } t\left(0, \sigma_t^2, \nu\right) \\ \sigma_t^2 \cdot Median\left[\chi_1^2\right] \approx 0.45\sigma_t^2, & \text{if } r_t|\mathcal{F}_{t-1} \sim N\left(0, \sigma_t^2\right) \end{cases}
\end{aligned} \quad (14)
$$

where $\varepsilon_t \equiv r_t/\sigma_t$ and $Median_{t-1}\left[r_t^2\right]$ is the conditional median of r_t^2 given $\mathcal{F}_{t-1}$. Thus, under normality, if we use MAE to compare a forecast which is exactly equal to σ_t^2 for all t to one that is equal to $0.45\sigma_t^2$ for all t, using the squared daily return as a proxy for the conditional variance, we will usually conclude that the perfect forecast is inferior to the one which is wrong by more than a factor of 2. Figure 1 shows that if returns have a Student's t distribution then the degree of distortion is even larger.

Another commonly used loss function is the MSE loss function on standard deviations rather than variances, see equation (8). The motivation for this loss function is that taking square root of the two arguments of the squared-error loss function shrinks the larger values towards zero, reducing the impact of the most extreme values of r_t. However it also leads to an incorrect volatility forecast being selected as optimal:

$$\begin{aligned}
h_t^* &\equiv \underset{h\in\mathcal{H}}{\arg\min}\; E_{t-1}\left[\left(|r_t| - \sqrt{h}\right)^2\right] \\
\text{FOC}\quad 0 &= \frac{\partial}{\partial h}\, E_{t-1}\left[\left(|r_t| - \sqrt{h}\right)^2\right]\Big|_{h=h_t^*} \\
\text{so}\quad h_t^* &= \left(E_{t-1}\left[|r_t|\right]\right)^2 \qquad (15) \\
&= \sigma_t^2 \left(E_{t-1}\left[|\varepsilon_t|\right]\right)^2 \\
&= \begin{cases} \frac{\nu-2}{\pi}\left(\Gamma\left(\frac{\nu-1}{2}\right)/\Gamma\left(\frac{\nu}{2}\right)\right)^2 \sigma_t^2, & \text{if } r_t|\mathcal{F}_{t-1} \sim \textit{Student's } t\left(0,\sigma_t^2,\nu\right),\ \nu>2 \\ \frac{2}{\pi}\sigma_t^2 \approx 0.64\sigma_t^2, & \text{if } r_t|\mathcal{F}_{t-1} \sim N\left(0,\sigma_t^2\right) \end{cases} \qquad (16)
\end{aligned}$$

For this loss function it is also true that excess kurtosis in asset returns exacerbates the distortion, which we can see in Figure 2 for returns that have the Student's t distribution.

In Appendix 1 we provide the corresponding calculations for the remaining loss functions in equations (5) to (13) above, and summarise the results in Table 1. Table 1 shows that the degree of distortion in the optimal forecast according to some of the loss functions used in the literature can be substantial. Under normality the optimal forecast under these loss functions ranges from about one quarter of the true conditional variance to three times the true conditional variance. If returns exhibit excess conditional kurtosis then the range of optimal forecasts from these loss functions is even wider.

Table 1 provides a theoretical explanation for the almost inevitable conflicting rankings of volatility forecasts that are obtained when non-robust loss functions are used in applied work. Lamoureux and Lastrapes (1993), Hamilton and Susmel (1994), Bollerslev and Ghysels (1996) and Hansen and Lunde (2005), amongst many others, use some or all of the nine loss functions considered in Table 1 and find that the best-performing volatility model changes with the choice

of loss function. Given that, for example, the MSE-prop loss function leads to an optimal forecast that is biased upwards by at least a factor of three, while the MAE loss function leads to an optimal forecast that is biased *downwards* by at least a factor of two, it is no surprise that different rankings of volatility forecasts are found.

To illustrate and emphasize the empirical relevance of the results of Table 1, consider the following example.

Example 1: Assume that $r_t|\mathcal{F}_{t-1} \sim N\left(0, \sigma_t^2\right)$, and that σ_t^2 follows a simple GARCH(1,1) process: $\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha r_{t-1}^2$, subject to $\omega > 0$ and $1 - \beta^2 - 2\alpha\beta - 3\alpha^2 > 0$ (which is required for $E\left[\sigma_t^4\right]$ to exist). Let $\hat{\sigma}_t^2 = r_t^2$, let L be the MSE-SD loss function, and let $h_{1t} = \sigma_t^2$ and $h_{2t} = 2/\pi\sigma_t^2$. Let n denote the number of observations available for conducting the test. Then the DMW test statistic evaluated at population moments is:

$$\begin{aligned} DMW_0 &= \left(\frac{5 + 3\sqrt{2/\pi}}{1 - \sqrt{2/\pi}} \cdot \frac{1 - (\alpha + \beta)^2}{1 - (\alpha + \beta)^2 - 2\alpha^2} - 1\right)^{-1/2} \cdot \sqrt{n} \\ &\approx 0.1632\sqrt{n}, \text{ when } \alpha = 0.05 \text{ and } \beta = 0.9. \end{aligned}$$

The derivation is in Appendix 1. For the specific case that $[\alpha, \beta] = [0.05, 0.9]$, which is reasonable for daily asset returns, the DMW_0 statistic is greater than 1.96 for sample sizes larger than 145. Thus with less than a year's worth of daily data, we would expect to reject the true conditional variance in favour of a volatility forecast equal to around 0.64 times the true conditional variance. This example shows that choosing an inappropriate loss function for volatility forecast comparison can have important empirical implications in realistic situations.

2.2 Using better volatility proxies

It has long been known that squared returns are a quite noisy proxy for the true conditional variance. One alternative volatility proxy that has gained much attention recently is "realised volatility", see Andersen, *et al.* (2001a, 2003), and Barndorff-Nielsen and Shephard (2002, 2004). Another commonly-used alternative to squared returns is the intra-daily range. It is well-known that if the log stock price follows a Brownian motion then both of these estimators are unbiased and more efficient than the squared return.

In this section we obtain the rate at which the distortion in the ranking of alternative forecasts disappears when using realised volatility as the proxy, as the sampling frequency increases, for a simple data generating process (DGP). These results can be viewed as complements to that of Hansen and Lunde (2006), who showed that under certain conditions the degree of distortion in

ranking alternative forecasts is increasing in the variability of the proxy error.

Assume that there are m equally-spaced observations per trade day, and let $r_{i,m,t}$ denote the i^{th} intra-daily return on day t. In order to obtain analytical results for problems involving the range as a volatility proxy we consider only a simple DGP: zero mean return, no jumps, and constant conditional volatility within a trade day[8]. Chen and Patton (2006) present corresponding results for a range of more realistic DGPs via simulation. Let

$$\begin{aligned} r_t &= d\log P_t = \sigma_t dW_t && (17)\\ \sigma_\tau &= \sigma_t \,\forall \tau \in (t-1, t] && (18)\\ r_{i,m,t} &\equiv \int_{(i-1)/m}^{i/m} r_\tau d\tau = \sigma_t \int_{(i-1)/m}^{i/m} dW_\tau && (19)\\ \text{so } \{r_{i,m,t}\}_{i=1}^{m} &\sim iid\, N\left(0, \frac{\sigma_t^2}{m}\right) && (20) \end{aligned}$$

We place no constraints on how σ_t^2 changes between trade days, though the assumption of constant intra-daily volatility is clearly restrictive. The "realised volatility" or "realised variance" is defined as:

$$RV_t \equiv \sum_{i=1}^{m} r_{i,m,t}^2$$

Realised variance, like the daily squared return (which is obtained in the above framework by setting $m = 1$), is a conditionally unbiased estimator of the daily conditional variance. Its main advantage is that it is more efficient estimator than the daily squared return: for this DGP it can be shown that $MSE_{t-1}\left[r_t^2\right] = 2\sigma_t^4$ while $MSE_{t-1}\left[RV_t\right] = 2\sigma_t^4/m$.

A volatility proxy that pre-dates realised volatility by many years is the range, or the high/low, estimator, see Parkinson (1980), Garman and Klass (1980) and Ball and Torous (1984). Alizadeh, *et al.* (2002) use the fact that the range is widely available and is more efficient than squared returns to improve the estimation of stochastic volatility models. The intra-daily log range is defined as:

$$RG_t \equiv \max_\tau \log P_\tau - \min_\tau \log P_\tau,\ t-1 < \tau \leq t \tag{21}$$

Under the dynamics in equation (17) Feller (1951) presented the density of RG_t, and Parkinson (1980) presented a formula for obtaining moments of the range, which enable us to compute:

$$E_{t-1}\left[RG_t^2\right] = 4\log(2)\cdot\sigma_t^2 \approx 2.7726\sigma_t^2 \tag{22}$$

[8] Analytical and empirical results on the range and "realised range" under more flexible DGPs are presented in two recent working papers by Christensen and Podolskij (2005) and Martens and van Dijk (2005).

Details on the distributional properties of the range under this DGP are presented in Appendix 1. The above expression shows that squared range is *not* a conditionally unbiased estimator of σ_t^2. Most authors, see Parkinson (1980) and Alizadeh, *et al.* (2002) for example, who employ the range as a volatility proxy are aware of this and scale the range accordingly. We will thus focus below on the *adjusted range*:

$$RG_t^* \equiv \frac{RG_t}{2\sqrt{\log(2)}} \approx 0.6006 RG_t \tag{23}$$

which, when squared, is an unbiased proxy for the conditional variance. Using the results of Parkinson (1980) it is simple to determine that $MSE_{t-1}\left[RG_t^{*2}\right] \approx 0.4073\sigma_t^4$, which is approximately one-fifth of the MSE of the daily squared return, and so using the range yields an estimator as accurate as a realised volatility estimator constructed using 5 intra-daily observations. This roughly corresponds to the comment of Andersen and Bollerslev (1998, footnote 20) that the adjusted range yields an MSE comparable to the MSE of realised volatilities constructed using 2 to 3 hour returns.

We now determine the optimal forecasts obtained using the various loss functions considered above, when $\hat{\sigma}_t^2 = RV_t$ or $\hat{\sigma}_t^2 = RG_t^{*2}$ is used as a proxy for the conditional variance rather than r_t^2. We initially leave m unspecified for the realised volatility proxy, and then specialise to three cases: $m = 1$, 13 and 78, corresponding to the use of daily, half-hourly and 5-minute returns, on a stock listed on the New York Stock Exchange (NYSE).

For MSE and QLIKE the optimal forecast is simply the conditional mean of $\hat{\sigma}_t^2$, which equals the conditional variance, as RV_t and RG_t^{*2} are both conditionally unbiased. The MSE-SD loss function yields $\left(E_{t-1}\left[\hat{\sigma}_t\right]\right)^2$ as an optimal forecast. Under the set-up introduced above,

$$\begin{aligned}
RV_t &\equiv \sum_{i=1}^{m} r_{t,i}^2 = \frac{\sigma_t^2}{m}\sum_{i=1}^{m}\varepsilon_{t,i}^2 \\
\text{so } m\sigma_t^{-2}RV_t &\sim \chi_m^2 \\
\text{so } h_t^* &= \frac{\sigma_t^2}{m}\left(E\left[\sqrt{\chi_m^2}\right]\right)^2 \\
E\left[\sqrt{\chi_m^2}\right] &\approx \sqrt{m} - \frac{1}{4\sqrt{m}} \text{ by a Taylor series approximation} \\
\text{so } h_t^* &\approx \sigma_t^2\left(1 - \frac{1}{2m} + \frac{1}{16m^2}\right) \\
&\approx \begin{cases} 0.5625\cdot\sigma_t^2 & \text{for } m = 1 \\ 0.9619\cdot\sigma_t^2 & \text{for } m = 13 \\ 0.9936\cdot\sigma_t^2 & \text{for } m = 78 \end{cases}
\end{aligned}$$

The results for the MSE-SD loss function using realised volatility show that reducing the noise

in the volatility proxy improves the optimal forecast, consistent with Hansen and Lunde (2006).[9] Using the range we find that

$$h_t^* = \left(E_{t-1}\left[RG_t^*\right]\right)^2 = \frac{2}{\pi \log 2}\sigma_t^2 \approx 0.9184\sigma_t^2$$

and so the distortion from using the range is approximately equal to that incurred when using a realised volatility constructed using 6 intra-daily observations.

Consider now the MAE loss function, which yields $Median_{t-1}\left[\hat{\sigma}_t^2\right]$ as the optimal forecast. For realised volatility we thus have

$$h_t^* = \frac{1}{m} Median\left[\chi_m^2\right]\sigma_t^2$$

For large m, $Median\left[\chi_m^2\right] \approx m - 2/3$, though most software packages have functions for the inverse *cdf* of a χ_m^2 distribution. For small m the approximation $Median\left[\chi_m^2\right] \approx m - 2/3 + 1/(9m)$ is more accurate. Thus

$$\begin{aligned} h_t^* &\approx \left(1 - \frac{2}{3m} + \frac{1}{9m^2}\right)\sigma_t^2 \\ &\approx \begin{cases} 0.4444 \cdot \sigma_t^2 & \text{for } m = 1 \\ 0.9494 \cdot \sigma_t^2 & \text{for } m = 13 \\ 0.9915 \cdot \sigma_t^2 & \text{for } m = 78 \end{cases} \quad \text{using } Median\left[\chi_m^2\right] \approx m - 2/3 + 1/(9m) \end{aligned}$$

For the range we have

$$h_t^* \approx \frac{2.2938}{\log 16}\sigma_t^2 = 0.8273\sigma_t^2$$

which is equivalent to using about 4 observations to construct the realised volatility proxy. Calculations for the remaining loss functions are collected in Appendix 1, and the results are summarised in Table 2.

The results in Table 2 confirm that as the proxy used to measure the true conditional variance gets more efficient the degree of distortion decreases for all loss functions. Across loss functions we found that the range was generally approximately as good a volatility proxy as the realised volatility estimator constructed with between 4 and 6 intra-daily observations. Using half-hour returns (13 intra-daily observations) or the intra-daily range still leaves substantial distortions in the optimal forecasts, but using 5-minute returns (78 intra-daily observations) eliminates almost all of the bias, at least in this simple framework[10].

[9] Note that the result for $m = 1$ is different to that obtained in Section 2, which was $h_t^* = \frac{2}{\pi}\sigma_t^2 \approx 0.6366\sigma_t^2$. This is because for $m = 1$ we can obtain the expression exactly, using results for the normal distribution, whereas for arbitrary m we relied on a second-order Taylor series approximation.

[10] Chen and Patton (2006) find very similar results to those in Table 2 when the DGP is specified to be log-normal,

2.3 General comments on non-robust loss functions

The sources of the mis-matches between the optimal forecast for a given loss function and the true conditional variance are easily identified. The MAE, MAE-SD and MAE-prop loss functions consider mean absolute distances rather than mean squared distances, which then naturally change the solution of the optimisation problem from an expectation to a median. For the MSE-log, MSE-SD and MSE-prop loss functions the distortion follows from the fact that the unbiasedness property is not invariant to nonlinear transformations.

In all of these cases the distortions can be remedied if one can obtain a conditionally unbiased estimator of the quantity of interest (σ_t, $\log \sigma_t^2$, etc.) either exactly or approximately. When using the squared return as a proxy, this will generally require an assumption about the entire conditional distribution of returns. When using realised variance as a volatility proxy one may obtain an approximate distribution of the volatility proxy under relatively mild assumptions, by drawing on the distribution theory for realised volatility developed in Barndorff-Nielsen and Shephard (2004) and extensions, as in Andersen, *et al.* (2005a). On the other hand, when using a robust loss function only the assumption of conditional unbiasedness of the proxy is required, which is often satisfied under much weaker assumptions and requires no adjustment of the proxy.

We now seek to identify the reason why some non-robust loss function yield upward-biased forecasts, whilst others yield downward-biased forecasts. We do so by generalising the results from the previous sections to a broad class of arbitrary loss functions, making use of Taylor series approximations. This requires some differentiability assumptions on the loss function, which are not satisfied for some of the loss functions considered above.

Assumption T1: The volatility proxy satisfies: $E_{t-1}\left[\hat{\sigma}_{t,m}^2\right] = \sigma_t^2$ and $V_{t-1}\left[\hat{\sigma}_{t,m}^2\right] = \nu_{t,m}^2$

Assumption T2: The loss function L is three times differentiable.

Assumption T3: The loss function L is such that $L\left(\sigma^2, h\right) = 0$ iff $h = \sigma^2$

Assumption T4: The loss function L is such that $\partial L\left(\sigma^2, h\right)/\partial h \gtreqless 0$ if $\sigma^2 \lesseqgtr h$

Assumption T5: The volatility proxy and the loss function are such that $L\left(\hat{\sigma}_{t,m}^2, h\right) - L\left(\sigma_t^2, h\right) \to^p 0$ as $m \to \infty$ uniformly on $\mathcal{H}$.

Proposition 1 *Define*

$$h_{t,m}^* \equiv \underset{h \in \mathcal{H}}{\arg\min}\ E_{t-1}\left[L\left(\hat{\sigma}_{t,m}^2, h\right)\right]$$

GARCH or two-factor stochastic volatility diffusions. Using the same parameterisations as those in the simulations of Goncalves and Meddahi (2005), they find slightly larger biases from the non-robust loss functions under these DGPs, but they generally differ from those in Table 2 only in the second decimal place.

(i) Let assumptions T1-T4 hold. Then

$$\frac{\partial^3 L\left(\sigma^2,h\right)}{\partial\left(\sigma^2\right)^2\partial h}\gtreqless 0 \text{ for all } \left(\sigma^2,h\right)\Rightarrow h^*_{t,m}\lesseqgtr\sigma^2_t$$

(ii) Let assumptions T3-T5 hold. Then $h^*_{t,m}\to^p\sigma^2_t$ *as* $m\to\infty$.

The first part of the above proposition shows that it is the sign of the third derivative of the loss function that determines whether the optimal forecast is above, below or equal to the true conditional variance. The case that this third derivative is equal to zero, and thus that the optimal forecast is the conditional variance, corresponds to a result of Hansen and Lunde (2006). The third derivative is always positive for the MSE-log and MSE-SD loss functions, and so part (i) above implies that $h^*_{t,m}<\sigma^2_t$ for these loss functions, which is consistent with the results in Tables 1 and 2. Alternatively, for the MSE-prop loss function this third derivative is always negative, implying that $h^*_{t,m}>\sigma^2_t$ for this loss function, which is again consistent with Tables 1 and 2.

The second part of the above proposition shows that under the high-level assumption of uniform convergence of $L\left(\hat{\sigma}^2_{t,m},h\right)$ to $L\left(\sigma^2_t,h\right)$, the optimal forecast converges to the conditional variance as $m\to\infty$. Thus even loss functions that cause distortions in the presence of noise in the volatility proxy can generate optimal forecasts that are consistent for the conditional variance, and so non-robust loss functions may be used in conjunction with proxies that can be assumed "nearly" perfect.

3 A class of robust loss functions

In the previous section we showed that amongst nine loss functions commonly used to compare volatility forecasts, only the MSE and the QLIKE loss functions lead to $h^*_t=E_{t-1}\left[\hat{\sigma}^2_t\right]=\sigma^2_t$, which is a necessary condition for a loss function to be robust to noise in the volatility proxy. The following proposition provides a necessary and sufficient class of robust loss functions, which are related to the class of linear-exponential densities of Gourieroux, *et al.* (1984), and to the work of Gourieroux, *et al.* (1987). We make the following assumptions:

A1: $E_{t-1}\left[\hat{\sigma}^2_t\right]=\sigma^2_t$

A2: $\hat{\sigma}^2_t|\mathcal{F}_{t-1}\sim F_t\in\tilde{F}$, the set of all absolutely continuous distribution functions on $\mathbb{R}_+$.

A3: L is twice continuously differentiable with respect to h and $\hat{\sigma}^2$, and has a unique minimum at $\hat{\sigma}^2=h$.

A4: There exists some $h^*_t\in int\left(\mathcal{H}\right)$ such that $h^*_t=E_{t-1}\left[\hat{\sigma}^2_t\right]$, where $\mathcal{H}$ is a compact subset of $\mathbb{R}_{++}$.

A5: L and F_t are such that: (a) $E_{t-1}\left[L\left(\hat{\sigma}_t^2, h\right)\right] < \infty$ for some $h \in \mathcal{H}$; (b) $\left|E_{t-1}\left[\partial L\left(\hat{\sigma}_t^2, \sigma_t^2\right)/\partial h\right]\right| < \infty$; and (c) $\left|E_{t-1}\left[\partial^2 L\left(\hat{\sigma}_t^2, \sigma_t^2\right)/\partial h^2\right]\right| < \infty$ for all t.

Proposition 2 *Let assumptions A1 to A5 hold. Then a loss function L is robust, in the sense of Definition 1, if and only if it takes the following form:*

$$L\left(\hat{\sigma}^2, h\right) = \tilde{C}\left(h\right) + B\left(\hat{\sigma}^2\right) + C\left(h\right)\left(\hat{\sigma}^2 - h\right) \tag{24}$$

where B and C are twice continuously differentiable, C is a strictly decreasing function on $\mathcal{H}$, and $\tilde{C}$ is the anti-derivative of C.

Remark 1 *If we normalise the loss function to yield zero loss when $\hat{\sigma}^2 = h$, then the class of robust loss functions takes the form:*

$$L\left(\hat{\sigma}^2, h\right) = \tilde{C}\left(h\right) - \tilde{C}\left(\hat{\sigma}^2\right) + C\left(h\right)\left(\hat{\sigma}^2 - h\right) \tag{25}$$

where C is a twice continuously differentiable, strictly decreasing function on $\mathcal{H}$, and $\tilde{C}$ is the anti-derivative of C.

Given the widespread interest in economics and finance in loss functions that depend only on the forecast error or the standardised forecast error, we present below a surprising result on the subset of robust loss functions that satisfy one of these restrictions.

Proposition 3 *(i) The "MSE" loss function is the only robust loss function that depends solely on the forecast error, $\hat{\sigma}^2 - h$.*

(ii) The "QLIKE" loss function is the only robust loss function that depends solely on the standardised forecast error, $\hat{\sigma}^2/h$.

The general representation of robust loss functions in Proposition 2 provides a simple means of determining whether a given loss function is suitable for use in volatility forecast comparison, but it does not directly provide new alternative robust loss functions. To this end, we now seek to find a parametric family of loss functions, that is a member of the class proposed above, and which nests MSE and QLIKE as special cases. We do this by noting that the first-order conditions from MSE and QLIKE loss functions are both of the form:

$$\frac{\partial L\left(\hat{\sigma}^2, h\right)}{\partial h} = 0 = ah^b\left(\hat{\sigma}^2 - h\right),\ a < 0,\ b \in \mathbb{R} \tag{26}$$

From this first-order condition we obtain the following parametric family of robust loss functions. Part (ii) below shows that this parametric family coincides with the subset of homogeneous robust loss functions.

Proposition 4 *(i) The following family of functions*

$$L\left(\hat{\sigma}^2, h; b\right) = \begin{cases} \frac{1}{(b+1)(b+2)}(\hat{\sigma}^{2b+4} - h^{b+2}) - \frac{1}{b+1}h^{b+1}\left(\hat{\sigma}^2 - h\right), & \text{for } b \notin \{-1, -2\} \\ h - \hat{\sigma}^2 + \hat{\sigma}^2 \log \frac{\hat{\sigma}^2}{h}, & \text{for } b = -1 \\ \frac{\hat{\sigma}^2}{h} - \log \frac{\hat{\sigma}^2}{h} - 1, & \text{for } b = -2 \end{cases} \tag{27}$$

satisfy $L(h, h; b) = 0$ *for all* $h \in \mathcal{H}$*, and are of the form in Proposition 2.*

(ii) The family of loss functions in part (i) corresponds to the entire subset of homogeneous robust loss functions. The degree of homogeneity is equal to $b + 2$*.*

The MSE loss function is obtained when $b = 0$ and the QLIKE loss function is obtained when $b = -2$, up to additive and multiplicative constants. In Figure 3 we present the above class of functions for various values of b, ranging from 1 to -5, and including the MSE and QLIKE cases. This figure shows that this family of loss functions can take a wide variety of shapes, ranging from symmetric ($b = 0$, corresponding to the MSE loss function) to asymmetric, with heavier penalty either on under-prediction ($b < 0$) or over-prediction ($b > 0$). Figure 4 plots the ratio of losses incurred for negative forecast errors to those incurred for positive forecast errors, to make clearer the form of asymmetries in these loss functions.

Given the arbitrariness of the choice of units in most economic and financial problems (for example, measuring prices in dollars versus cents, or measuring returns in percentages versus decimals) it is potentially interesting to consider the impact of a simple change in units on the ranking of two competing forecasts by expected loss. The class of loss functions presented in Proposition 2 guarantees that the true conditional variance will be chosen (subject to sampling variation) over any other forecast regardless of the choice units. However it does *not* guarantee that the ranking of two *imperfect* forecasts will be invariant to the choice of units. The following proposition shows that by using a homogeneous robust loss function, as in Proposition 4, the ranking of any two (possibly imperfect) forecasts is invariant to a re-scaling of the data. It further provides an example where the ranking can be reversed simply with a re-scaling of the data if a non-homogeneous robust loss function is used.

Proposition 5 *(i) The ranking of any two (possibly imperfect) volatility forecasts by expected loss is invariant to a re-scaling of the data if the loss function is robust and homogeneous.*

(ii) The ranking of any two (possibly imperfect) volatility forecasts by expected loss may not be invariant to a re-scaling of the data if the loss function is robust but not homogeneous.

Having presented a new class of loss functions, it is next of interest to establish the conditions under which we can employ these loss functions in DMW tests for volatility forecast comparison.

The main conditions to be determined are moment conditions on the volatility proxy and volatility forecasts, and these are presented in part (ii) of the following proposition.

Proposition 6 *Let* $d_t(b) \equiv L\left(\hat{\sigma}_t^2, h_{1t}; b\right) - L\left(\hat{\sigma}_t^2, h_{2t}; b\right)$. *(i) For a given loss function parameter* b, *and given that*

1. *(a)* $d_t(b) = d^0(b) + \varepsilon_t(b)$, $t = 1, 2, ...$; $d^0(b) \in \mathbb{R}$,

   (b) $\{d_t(b)\}$ *is a mixing sequence with either* ϕ *of size* $-r/2(r-1)$ *for some* $r \geq 2$, *or* α *of size* $-r/(r-2)$ *for some* $r > 2$,

   (c) $E[d_t(b)] = d^0(b)$ *for* $t = 1, 2, ...$,

   (d) $E[|d_t(b)|^r] < \Delta < \infty$ *for all* t, *and*

   (e) $V_n(b) \equiv V\left[n^{-1/2}\sum_{t=1}^n \varepsilon_t(b)\right]$ *is uniformly positive definite.*

Then

$$\frac{\sqrt{n}\left(\bar{d}(b) - d^0(b)\right)}{\sqrt{V_n(b)}} \to^{\mathcal{D}} N(0,1), \text{ as } n \to \infty$$

where $\bar{d}_n(b) \equiv n^{-1}\sum_{t=1}^n d_t(b)$. *Under* $H_0 : E[d_t(b)] = 0$, *we have:*

$$DMW_n(b) \equiv \frac{\sqrt{n}\bar{d}_n(b)}{\sqrt{\hat{V}\left[\sqrt{n}\bar{d}_n(b)\right]}} \to^{\mathcal{D}} N(0,1) \text{ as } n \to \infty$$

where $\hat{V}\left[\sqrt{n}\bar{d}_n(b)\right]$ *is any consistent estimator of* $V\left[\sqrt{n}\bar{d}_n(b)\right]$. *If* $E[d_t(b)] \neq 0$ *then* $DMW_n(b) \to \pm\infty$.

(ii) Sufficient conditions for $E\left[d_t(b)^2\right] < \infty$ *are*

1. $\inf_t h_{it} \equiv c_i > 0$ *for* $i = 1, 2$,

2. $E[h_{it}^p] < \infty$, $i = 1, 2$, *and*

3. $E[\hat{\sigma}_t^q] < \infty$,

where p *and* q *are as follows:*

$$\begin{aligned}
&p = \max[0, 2b+4], \ q = \max[4+\delta, 4b+8], \text{ for } \delta > 0, && \text{when } b \notin \{-1, -2\} \\
&p = 2(e+1)/e \approx 2.74, \ q = 4(e+1)/e \approx 5.47, && \text{when } b = -1 \\
&p = 2/e + \delta \approx 0.74 + \delta, \ q = 4 + \delta, \text{ for } \delta > 0, && \text{when } b = -2
\end{aligned}$$

where e *is the exponential constant,* $e \approx 2.71$.

The assumption that the volatility forecasts will never be less than some positive threshold is true for many standard volatility models, such as the GARCH(1,1), for example. Part (ii) of the above proposition show how greatly the moment conditions can vary depending on the choice of loss function shape parameter b. For MSE loss, corresponding to $b = 0$, we need $E\left[h_{it}^4\right]$ and $E\left[\hat{\sigma}_t^8\right]$ to be finite, whereas for the QLIKE loss function we only require $E\left[h_{it}^{2/e+\delta}\right]$ and $E\left[\hat{\sigma}_t^{4+\delta}\right]$, for $\delta > 0$, to be finite. Choosing $b \leq -2$ is recommended if the existence of moments of the volatility proxy or volatility forecasts is a concern.

4 Empirical application to forecasting IBM return volatility

In this section we consider the problem of forecasting the conditional variance of the daily return on IBM, using data from the TAQ database over the period from January 1993 to December 2003. We consider two simple volatility models that are widely-used in industry: a 60-day rolling window estimator, and the RiskMetrics volatility model based on daily returns:

$$\text{Rolling window} \quad : \quad h_{1t} = \frac{1}{60}\sum\nolimits_{j=1}^{60} r_{t-j}^2 \tag{28}$$

$$\text{RiskMetrics} \quad : \quad h_{2t} = \lambda h_{2t-1} + (1-\lambda)\, r_{t-1}^2,\ \lambda = 0.94 \tag{29}$$

We use approximately the first year of observations (272 observations) to initiate the RiskMetrics forecasts, and the remaining 2500 observations to compare the forecasts. A plot of the volatility forecasts is provided in Figure 5.

We employ a variety of volatility proxies in the comparison of these forecasts: the daily squared return, and realised variance computed using 65-minute, 15-minute and 5-minute returns[11]. In comparing these forecasts we present the results of Diebold-Mariano-West tests using the loss function presented in Proposition 4, for five different choices of the loss function parameter: $b = \{1, 0, -1, -2, -5\}$. MSE loss and QLIKE loss correspond to $b = 0$ and $b = -2$ respectively. Recall from the previous section that different choices of b require weaker or stronger moment conditions for the DMW test to be valid. For $b = -5$ we only require $E\left[\hat{\sigma}_t^{4+\delta}\right] < \infty$ for $\delta > 0$, whereas for $b = 1$ we need $E\left[h_{it}^6\right] < \infty$ and $E\left[\hat{\sigma}_t^{12}\right] < \infty$. These assumptions should be kept in mind when interpreting the results below.

Table 3 presents the results of standard Mincer-Zarnowitz tests of the volatility forecasts. Both the rolling window and the RiskMetrics forecasts are rejected using all four volatility proxies, with

[11] We use 65-minute returns rather than 60-minute returns so that there are an even number of intervals within the NYSE trade day, which runs from 9.30am to 4pm.

MZ test p-values equal to 0.00 in all cases. We can thus conclude that neither of these forecasts is optimal. This conclusion leads then to the question of relative forecast performance, for which we use a DMW test.

In Table 4 we present tests comparing the RiskMetrics forecasts based on daily returns with the 60-day rolling window volatility forecasts. The only loss function for which the difference in forecast performance is significantly different from zero is the QLIKE loss function: the difference is significant at the 0.05 level using 65-minute, 15-minute and 5-minute realised variances as the volatility proxy, and significant at the 0.10 level using daily squared returns as the proxy. In all of these cases the t-statistic is positive, indicating that the rolling window forecasts generated larger average loss than the RiskMetrics forecasts. Interestingly, under MSE loss, the differences in average loss favour the rolling window forecasts, though these differences are not statistically significant.

5 Conclusion

We analytically demonstrated some problems with volatility forecast comparison techniques used in the literature. These techniques invariably rely on a volatility proxy, which is some imperfect estimator of the true conditional variance, and the presence of noise in the volatility proxy can lead an imperfect volatility forecast being selected over the true conditional variance for certain choices of loss function. We showed analytically that less noisy volatility proxies, such as the intra-daily range and realised volatility, lead to less distortion, though in some cases the degree of distortion is still large.

We derived necessary and sufficient conditions on the loss function for it to yield rankings of volatility forecasts that are robust to noise in the proxy. We also proposed a new parametric family of robust loss functions and derived the moment conditions necessary for the use of this loss function in forecast comparison tests. The new family of loss function nests both squared-error and the "QLIKE" loss functions, two of the most widely-used in the volatility forecasting literature. A small empirical study of IBM equity volatility illustrated the new loss functions in forecast comparison tests.

Whilst volatility forecasting is a prominent example of a problem in economics where the variable of interest is unobserved, there are many other such examples: forecasting the true rates of inflation or GDP growth (not simply the announced rates); forecasting trade intensities; forecasting default probabilities or 'crash' probabilities; and forecasting covariances or correlations. The derivations in

this paper exploited the fact that the latent variable of interest in volatility forecasting (namely the conditional variance) is a positive random variable, and the proxy is non-negative and continuously distributed. Extending the results in this paper to handle latent variables of interest with support on the entire real line, as would be required for applications to studies of the "true" rates of growth in macroeconomic aggregates or to conditional covariances, should not be difficult. Extending our results to handle proxies with discrete support, such as those that would be used in default forecasting applications, may require a different method of proof. We leave such extensions to future research.

6 Appendix 1: Supporting calculations for Section 2

Section 2.1:

Optimal forecasts under alternative loss functions. Recall that $\varepsilon_t \equiv r_t/\sigma_t$.

MSE-log:

$$\begin{aligned} h_t^* &= \exp\left\{E_{t-1}\left[\log \varepsilon_t^2\right]\right\}\sigma_t^2 \\ &= \begin{cases} (\nu-2)\exp\left\{\Psi\left(\frac{1}{2}\right)-\Psi\left(\frac{\nu}{2}\right)\right\}\sigma_t^2, & \text{if } r_t|\mathcal{F}_{t-1} \sim \text{Student's } t\left(0,\sigma_t^2,\nu\right),\ \nu>2 \\ \frac{1}{2}\exp\left\{-\gamma_E\right\}\sigma_t^2 \ \approx 0.28\sigma_t^2, & \text{if } r_t|\mathcal{F}_{t-1} \sim N\left(0,\sigma_t^2\right) \end{cases} \end{aligned}$$

where Ψ is the digamma function and $\gamma_E = -\Psi(1) \approx 0.58$ is Euler's constant, see Harvey, et al. (1994).

MSE-prop:

$$\begin{aligned} h_t^* &= \frac{E_{t-1}\left[r_t^4\right]}{E_{t-1}\left[r_t^2\right]} = Kurtosis_{t-1}\left[r_t\right]\sigma_t^2 \\ &= \begin{cases} 3\left(\frac{\nu-2}{\nu-4}\right)\sigma_t^2, & \text{if } r_t|\mathcal{F}_{t-1} \sim \text{Student's } t\left(0,\sigma_t^2,\nu\right),\ \nu>4 \\ 3\sigma_t^2, & \text{if } r_t|\mathcal{F}_{t-1} \sim N\left(0,\sigma_t^2\right) \end{cases} \end{aligned}$$

QLIKE: $h_t^* = E_{t-1}\left[r_t^2\right] = \sigma_t^2$

MAE-log: $h_t^* = \exp\left\{Median_{t-1}\left[\log \varepsilon_t^2\right]\right\}\sigma_t^2 = Median_{t-1}\left[\varepsilon_t^2\right]\sigma_t^2$, since $Median\left[\log(X)\right] = \log\left(Median\left[X\right]\right)$ for any non-negative random variable X. Thus the optimal forecast is identical to that under MAE loss, which is given in the body of the paper.

MAE-SD: $h_t^* = \left(Median_{t-1}\left[|\varepsilon_t|\right]\right)^2\sigma_t^2 = Median_{t-1}\left[\varepsilon_t^2\right]\sigma_t^2$, since $Median\left[X\right]^2 = Median\left[X^2\right]$ for any non-negative random variable X. Thus the optimal forecast is identical to that under MAE loss, which is given in the body of the paper.

MAE-prop: If $r_t^2|\mathcal{F}_{t-1} \sim F_t\left(\sigma_t^2\right)$ and $\varepsilon_t^2 \equiv r_t^2/\sigma_t^2|\mathcal{F}_{t-1} \sim G_t\left(1\right)$ then

$$\begin{aligned} FOC \quad 0 &= \int_0^{h_t^*} \frac{r_t^2}{h_t^*} f_t\left(r_t^2\right) dr_t^2 - \int_{h_t^*}^{\infty} \frac{r_t^2}{h_t^*} f_t\left(r_t^2\right) dr_t^2 \\ \text{so} \quad \int_0^{h_t^*} \frac{r_t^2}{h_t^*} f_t\left(r_t^2\right) dr_t^2 &= \int_{h_t^*}^{\infty} \frac{r_t^2}{h_t^*} f_t\left(r_t^2\right) dr_t^2 \\ F_t\left(h_t^*\right) E_{t-1}\left[\frac{r_t^2}{h_t^*}\middle| r_t^2 \leq h_t^*\right] &= \left(1 - F_t\left(h_t^*\right)\right) E_{t-1}\left[\frac{r_t^2}{h_t^*}\middle| r_t^2 > h_t^*\right] \end{aligned}$$

without loss of generality let $h_t^* \equiv \sigma_t^2 \gamma_t^*$, $\gamma_t^* > 0$, so

$$\begin{aligned} F_t\left(\sigma_t^2\gamma_t^*\right) E_{t-1}\left[\frac{\varepsilon_t^2}{\gamma_t^*}\middle| \varepsilon_t^2 \leq \gamma_t^*\right] &= \left(1 - F_t\left(\sigma_t^2\gamma_t^*\right)\right) E_{t-1}\left[\frac{\varepsilon_t^2}{\gamma_t^*}\middle| \varepsilon_t^2 > \gamma_t^*\right] \\ G_t\left(\gamma_t^*\right) E_{t-1}\left[\varepsilon_t^2\middle| \varepsilon_t^2 \leq \gamma_t^*\right] &= \left(1 - G_t\left(\gamma_t^*\right)\right) E_{t-1}\left[\varepsilon_t^2\middle| \varepsilon_t^2 > \gamma_t^*\right] \end{aligned}$$

If $\varepsilon_t^2|\mathcal{F}_{t-1} \sim G\left(1\right)$, then $\gamma_t^* = \gamma^* \ \forall \ t$. Finding an explicit expression for h_t^* is difficult, and so we used 10,000 simulated draws for $\nu = \{4, 6, 10, 20, 30, 50, 100, 1000, \infty\}$ and numerically obtained h_t^* for each ν. We then used OLS to find the approximation given in Table 1, which yielded an R^2 of 0.9667.

DMW test using MSE-SD loss: We have

$$d_t = \left(|r_t| - \sigma_t\right)^2 - \left(|r_t| - \sqrt{2/\pi}\sigma_t\right)^2$$

and we seek to find an expression for DMW_0 as a function of $(\omega, \alpha, \beta, n)$, where $DMW_0 \equiv V\left[\sqrt{n}\bar{d}_n\right]^{-1/2} \sqrt{n} E\left[d_t\right]$. In the interests of parsimony we present results under the incorrect assumption that d_t is serially uncorrelated, which leads to the simplification $DMW_0 = V\left[d_t\right]^{-1/2} \sqrt{n} E\left[d_t\right]$. In unreported work we also derived the variance allowing for serial correlation in d_t and found that accounting for the serial correlation does not change the conclusion significantly. The serial correlation in d_t turns out to be negative, and so the correct variance is slightly smaller than the naïve variance estimator used, which makes the coefficient on $\sqrt{n}$ even larger.

$$\begin{aligned} d_t &= \left(|r_t| - \sigma_t\right)^2 - \left(|r_t| - \sqrt{2/\pi}\sigma_t\right)^2 = \left(1 - 2/\pi\right)\sigma_t^2 + 2\left(\sqrt{2/\pi} - 1\right)|\varepsilon_t|\,\sigma_t^2 \\ \text{so} \quad E\left[d_t\right] &= \left(1 - \sqrt{2/\pi}\right)^2 E\left[\sigma_t^2\right] \end{aligned}$$

and

$$\begin{aligned} E\left[d_t^2\right] &= E\left[\sigma_t^4 E_{t-1}\left[\left(1 - 2/\pi + 2\left|\varepsilon_t\right|\left(\sqrt{2/\pi} - 1\right)\right)^2\right]\right] \\ &= \left(1 - \sqrt{2/\pi}\right)^3 \left(5 + 3\sqrt{2/\pi}\right) E\left[\sigma_t^4\right] \end{aligned}$$

The quantities $E\left[\sigma_t^2\right]$ and $E\left[\sigma_t^4\right]$ depend on the DGP for the returns, and in this case they equal:

$$\begin{aligned} E\left[\sigma_t^2\right] &= \frac{\omega}{1-\alpha-\beta}, \text{ if } \alpha+\beta<1 \\ E\left[\sigma_t^4\right] &= \frac{\omega^2\left(1+\alpha+\beta\right)}{\left(1-\left(\alpha+\beta\right)^2-2\alpha^2\right)\left(1-\alpha-\beta\right)}, \text{ if } \left(1-\left(\alpha+\beta\right)^2-2\alpha^2\right)>0 \end{aligned}$$

so

$$\begin{aligned} DMW_0 &= \frac{\sqrt{n}\left(1-\sqrt{2/\pi}\right)^2 E\left[\sigma_t^2\right]}{\sqrt{\left(1-\sqrt{2/\pi}\right)^3\left(5+3\sqrt{2/\pi}\right)E\left[\sigma_t^4\right]-\left(1-\sqrt{2/\pi}\right)^4 E\left[\sigma_t^2\right]^2}} \\ &= \left(\frac{5+3\sqrt{2/\pi}}{1-\sqrt{2/\pi}}\frac{1-\left(\alpha+\beta\right)^2}{1-\left(\alpha+\beta\right)^2-2\alpha^2}-1\right)^{-1/2}\sqrt{n} \end{aligned}$$

as stated in the text. Note that the parameter ω does not affect the statistic.

Section 2.2:

Wherever possible we derived solutions or approximate solutions analytically. This was not always possible and so in some cases we had to resort to simulations to obtain solutions. Feller (1951) presents the density of the range:

$$f\left(RG_t;\sigma_t\right)=8\sum_{k=1}^{\infty}\left(-1\right)^{k-1}\frac{k^2}{\sigma_t}\phi\left(\frac{k\cdot RG_t}{\sigma_t}\right)$$

where ϕ is the standard normal *pdf*. For practical purposes the sum in the above expression needs to be truncated at some finite value; we truncate at $k=1000$. Parkinson (1980) presented the *cdf* of the range, and a formula for obtaining moments:

$$\begin{aligned} F\left(RG_t;\sigma_t\right) &= \sum_{k=1}^{\infty}\left(-1\right)^{k-1}k\left\{\text{erfc}\left(\frac{\left(k+1\right)RG_t}{\sigma\sqrt{2}}\right)-2\text{erfc}\left(\frac{k\cdot RG_t}{\sigma\sqrt{2}}\right)+\text{erfc}\left(\frac{\left(k-1\right)RG_t}{\sigma\sqrt{2}}\right)\right\} \\ E\left[RG_t^p\right] &= \frac{4}{\sqrt{\pi}}\Gamma\left(\frac{p+1}{2}\right)\left(2^{p/2}-2^{2-p/2}\right)\zeta\left(p-1\right)\sigma_t^p, \text{ for } p\geq 1 \end{aligned}$$

where $\text{erfc}(x)\equiv 1-\text{erf}(x)$, $\text{erf}(x)$ is the 'error function': $\text{erf}\left(x\right)\equiv 2/\sqrt{\pi}\int_0^x e^{-t^2}dt$. ζ is the Riemann zeta function. From this expression we can obtain the necessary moments for computing optimal forecasts when the range is used as a volatility proxy. For the first and second moments of RG_t we can obtain simple expressions, but the fourth moment involves $\zeta\left(3\right)=\Sigma_{k=1}^{\infty}k^{-3}$ which is an irrational number, and thus only a numerical expression is available. In addition to the moments of RG_t, we will need the mean of $\log RG_t$ and the median of RG_t. We used quadrature and OLS to obtain the expression[12]:

$$E_{t-1}\left[\log RG_t\right]=0.4257+\log\sigma_t \tag{30}$$

[12] We used quadrature to estimate $E_{t-1}\left[\log RG_t\right]$ for $\sigma_t=0.5,1,1.5,...,10$. We then regressed these estimates on a constant and $\log\sigma_t$ to obtain the parameter estimates. The R^2 from this regression was 1.0000.

which is consistent with the expression given in Alizadeh, et al. (2002). We numerically inverted the *cdf* of the range, given in Parkinson (1980), and used OLS to determine the following relation[13]:

$$\begin{aligned} Median_{t-1}\left[RG_t\right] &= 1.5145\sigma_t \\ \text{so } Median_{t-1}\left[RG_t^2\right] &= 2.2938\sigma_t^2, \text{ since } RG_t \text{ is weakly positive.} \end{aligned}$$

MSE-LOG: $h_t^* = \exp\left\{E_{t-1}\left[\log\hat{\sigma}_t^2\right]\right\}$. A Taylor series approximation did not provide a good fit when considering realised variance as a proxy, and so we resorted to simulations. We simulated 50,000 "days" worth of observations, where the number of observations per day considered was $m = \{1, 3, 5, 7, 10, 13, 20, 40, 60, 78, 100\}$. The following expression yielded an R^2 of 0.9959 : $E_{t-1}\left[\log RV_t^{(m)}\right] \approx -1.2741/m$, so the optimal forecast under our DGP assumption is $h_t^* \approx \sigma_t^2 e^{-1.2741/m}$.

For the range we find that

$$\begin{aligned} E_{t-1}\left[\log RG_t^{*2}\right] &= 2E_{t-1}\left[\log RG_t^*\right] \\ &= -0.1684 + \log\sigma_t^2 \\ \text{so } h_t^* &= e^{-0.1684}\sigma_t^2 \approx 0.8450\sigma_t^2 \end{aligned}$$

MAE-log: The optimal forecast is $h_t^* = Median_{t-1}\left[\hat{\sigma}_t^2\right]$, since $\hat{\sigma}_t^2$ is weakly positive we know that $\log\left(Median_{t-1}\left[\hat{\sigma}_t^2\right]\right) = Median_{t-1}\left[\log\hat{\sigma}_t^2\right]$, and so the results for this loss function are identical to those for the MAE loss function.

MAE-SD: The optimal forecast is $h_t^* = Median_{t-1}\left[\hat{\sigma}_t^2\right]$. Since $\hat{\sigma}_t^2$ is weakly positive we know that $Median_{t-1}\left[\hat{\sigma}_t^2\right] = \left(Median_{t-1}\left[\hat{\sigma}_t\right]\right)^2$, and so the results for this loss function are identical to those for the MAE loss function.

MSE-prop: $h_t^* = E_{t-1}\left[\hat{\sigma}_t^4\right]/E_{t-1}\left[\hat{\sigma}_t^2\right]$. When realised volatility is used as the proxy we find: $h_t^* = \left(\frac{1}{m}Kurtosis_{t-1}\left[r_{t,i}\right] + \frac{m-1}{m}\right)\sigma_t^2 = \left(1 + \frac{2}{m}\right)\sigma_t^2$. For the range we find that: $h_t^* = 10.8185/\left(\left(\log 16\right)^2\right)\sigma_t^2 \approx 1.4073\sigma_t^2$.

MAE-prop: For realised variance, like the daily squared return, obtaining an analytical, even approximate, solution to this problem is difficult and so we used simulations. In the set-up given in the text it is again possible to show that the optimal forecast is of the form $h_t^* = \gamma^*\sigma_t^2$. For realised volatility we simulated 50,000 "days" worth of observations, where the number of observations per day considered was

[13] The R^2 from this relation for $\sigma = 0.5, 1, 1.5, ..., 10$ was 1.0000.

$m = \{1, 3, 5, 7, 10, 13, 20, 40, 60, 78, 100\}$, and used numerical methods to locate the optimum forecast. The following expression yielded an R^2 of 0.9999 : $h_t^* \approx \left(1 + \frac{1.3624}{m}\right)\sigma_t^2$. For the range we again used a numerical minimisation algorithm combined with quadrature to compute the expectation in the optimisation problem: $h_t^* \approx 0.9941\sigma_t^2$.

7 Appendix 2: Proofs of Propositions

Proof of Proposition 1. (i) Approximate the loss function L with a second-order Taylor series:

$$\begin{aligned} L\left(\hat{\sigma}_{t,m}^2, h\right) &\approx L\left(\sigma_t^2, h\right) + \frac{\partial L\left(\sigma_t^2, h\right)}{\partial \sigma_t^2}\left(\hat{\sigma}_{t,m}^2 - \sigma_t^2\right) + \frac{1}{2}\frac{\partial^2 L\left(\sigma_t^2, h\right)}{\partial\left(\sigma_t^2\right)^2}\left(\hat{\sigma}_{t,m}^2 - \sigma_t^2\right)^2 \\ \text{so } E_{t-1}\left[L\left(\hat{\sigma}_{t,m}^2, h\right)\right] &\approx L\left(\sigma_t^2, h\right) + \frac{1}{2m}\frac{\partial^2 L\left(\sigma_t^2, h\right)}{\partial\left(\sigma_t^2\right)^2}\nu_{t,m}^2 \end{aligned}$$

by assumption T1. The first-order condition for forecast optimality is

$$\begin{aligned} 0 &= E_{t-1}\left[\frac{\partial L\left(\hat{\sigma}_{t,m}^2, h_{t,m}^*\right)}{\partial h}\right] \\ &\approx \frac{\partial L\left(\sigma_t^2, h_{t,m}^*\right)}{\partial h} + \frac{1}{2m}\frac{\partial^3 L\left(\sigma_t^2, h_{t,m}^*\right)}{\partial\left(\sigma_t^2\right)^2 \partial h}\nu_{t,m}^2 \end{aligned}$$

In the absence of noise in the volatility proxy (i.e. $\nu_{t,m}^2 = 0$) the second term above would equal zero and the first-order condition would be the same as if the true conditional variance was observable. By assumption T4 this yields $h_{t,m}^* = \sigma_t^2$. One of the conditions of Hansen and Lunde (2006) was $\partial^3 L\left(\sigma^2, h\right)/\partial\left(\sigma^2\right)^2 \partial h = 0$, which implies that the second term above equals zero even in the presence of a noisy volatility proxy. For loss functions that yield $\partial^3 L\left(\sigma^2, h\right)/\partial\left(\sigma^2\right)^2 \partial h \neq 0$ the presence of noise in the volatility proxy distorts the first-order condition from what it would be in the absence of noise, and thus affects the optimal forecast. If $\partial^3 L\left(\sigma^2, h\right)/\partial\left(\sigma^2\right)^2 \partial h > (<)\ 0\ \forall\ \left(\sigma^2, h\right)$, then the FOC implies that we must have $\partial L\left(\sigma_t^2, h_{t,m}^*\right)/\partial h < (>)\ 0$, which implies that $h_{t,m}^* < (>)\ \sigma_t^2$, by assumption T4.

(ii) Follows from Theorem 3.4 of White (1994), noting that assumptions T3 and T4 imply that $h^* = \sigma^2$ is the unique solution to the problem $\min_{h \in \mathcal{H}} L\left(\sigma^2, h\right)$. ■

Proof of Proposition 2. We prove this proposition by showing the equivalence of the following three statements:

$\mathcal{S}1$: The loss function takes the form given the statement of the proposition;

$\mathcal{S}2$: The loss function is robust in the sense of Definition 1;

$\mathcal{S}3$: The optimal forecast under the loss function is the conditional variance.

We will show that $\mathcal{S}1 \Rightarrow \mathcal{S}2$, and then that $\mathcal{S}1 \Leftrightarrow \mathcal{S}3$, and finally that $\mathcal{S}2 \Rightarrow \mathcal{S}3$.

That $\mathcal{S}1 \Rightarrow \mathcal{S}2$ follows from Hansen and Lunde (2006): their assumption 2 is satisfied given the assumptions for the proposition and noting that $\partial^2 L\left(\hat{\sigma}^2, h\right)/\partial\left(\hat{\sigma}^2\right)^2 = B''\left(\hat{\sigma}^2\right)$ does not depend on h.

We next show that $\mathcal{S}1 \Rightarrow \mathcal{S}3$: The first-order condition defining the optimal forecast is:

$$\begin{aligned} 0 &= \frac{\partial}{\partial h}\left(E_{t-1}\left[L\left(\hat{\sigma}_t^2, h_t^*\right)\right]\right) \\ &= \frac{\partial}{\partial h}\left(\tilde{C}\left(h_t^*\right) + E_{t-1}\left[B\left(\hat{\sigma}_t^2\right)\right] + C\left(h_t^*\right)\left(E_{t-1}\left[\hat{\sigma}_t^2\right] - h_t^*\right)\right) \\ &= C'\left(h_t^*\right)\left(E_{t-1}\left[\hat{\sigma}_t^2\right] - h_t^*\right) \end{aligned}$$

which implies $h_t^* = E_{t-1}\left[\hat{\sigma}_t^2\right]$ since C is a strictly decreasing function. The second-order condition is also satisfied: $\partial^2\left(E_{t-1}\left[L\left(\hat{\sigma}_t^2, h_t^*\right)\right]\right)/\partial h^2 = C''\left(h_t^*\right)\left(E_{t-1}\left[\hat{\sigma}_t^2\right] - h_t^*\right) - C'\left(h_t^*\right) = -C'\left(h_t^*\right) > 0$, since $h_t^* = E_{t-1}\left[\hat{\sigma}_t^2\right]$ and C is strictly decreasing.

Proving $\mathcal{S}3 \Rightarrow \mathcal{S}1$ is more challenging. For this part we follow the proof of Theorem 1 of Komunjer and Vuong (2004), adapted to our problem. We seek to show that the functional form of the loss function given in the proposition is necessary for $h_t^* = E_{t-1}\left[\hat{\sigma}_t^2\right]$, for any $F_t \in \tilde{F}$. Notice that we can write

$$\frac{\partial L\left(\hat{\sigma}_t^2, h_t\right)}{\partial h} = c\left(\hat{\sigma}_t^2, h_t\right)\left(\hat{\sigma}_t^2 - h_t\right)$$

where $c\left(\hat{\sigma}_t^2, h_t\right) = \left(\hat{\sigma}_t^2 - h_t\right)^{-1}\partial L\left(\hat{\sigma}_t^2, h_t\right)/\partial h$, since $\hat{\sigma}_t^2 \neq h_t$ a.s. by assumption A2. Now decompose $c\left(\hat{\sigma}_t^2, h_t\right)$ into

$$c\left(\hat{\sigma}_t^2, h_t\right) = E_{t-1}\left[c\left(\hat{\sigma}_t^2, h_t\right)\right] + \varepsilon_t$$

where $E_{t-1}\left[\varepsilon_t\right] = 0$. Thus

$$\begin{aligned} E_{t-1}\left[\frac{\partial L\left(\hat{\sigma}_t^2, h_t^*\right)}{\partial h}\right] &= E_{t-1}\left[c\left(\hat{\sigma}_t^2, h_t^*\right)\left(\hat{\sigma}_t^2 - h_t^*\right)\right] \\ &= E_{t-1}\left[c\left(\hat{\sigma}_t^2, h_t\right)\right] E_{t-1}\left[\hat{\sigma}_t^2 - h_t^*\right] + E_{t-1}\left[\varepsilon_t\left(\hat{\sigma}_t^2 - h_t^*\right)\right] \end{aligned}$$

If $E_{t-1}\left[\partial L\left(\hat{\sigma}_t^2, h_t^*\right)/\partial h\right] = 0$ for $h_t^* = E_{t-1}\left[\hat{\sigma}_t^2\right]$, then it must be that $E_{t-1}\left[\hat{\sigma}_t^2 - h_t^*\right] = 0 \Rightarrow E_{t-1}\left[\varepsilon_t\left(\hat{\sigma}_t^2 - h_t^*\right)\right] = 0$ for all $F_t \in \tilde{F}$. Employing a generalised Farkas lemma, see Lemma 8.1 of Gourieroux and Monfort (1996), this implies that $\exists\ \lambda \in \mathbb{R}$ such that $\lambda\left(\hat{\sigma}_t^2 - h_t^*\right) = \varepsilon_t\left(\hat{\sigma}_t^2 - h_t^*\right)$ for every $F_t \in \tilde{F}$ and for all t. Since $\hat{\sigma}_t^2 - h_t^* \neq 0$ a.s. by assumption A2 this implies that $\varepsilon_t = \lambda$ a.s. for all t. Since $E_{t-1}\left[\varepsilon_t\right] = 0$ we then have $\lambda = 0$. Thus $c\left(\hat{\sigma}_t^2, h_t^*\right) = E_{t-1}\left[c\left(\hat{\sigma}_t^2, h_t^*\right)\right]$ for all t, which implies that $c\left(\hat{\sigma}_t^2, h_t^*\right) = c\left(h_t^*\right)$, and thus that $\partial L\left(\hat{\sigma}_t^2, h_t\right)/\partial h = c\left(h_t\right)\left(\hat{\sigma}_t^2 - h_t\right)$.

A necessary condition for h_t^* to minimise $E_{t-1}\left[L\left(\hat{\sigma}_t^2, h\right)\right]$ is that $E_{t-1}\left[\partial^2 L\left(\hat{\sigma}_t^2, h_t^*\right)/\partial h^2\right] \geq 0$, using A5 to interchange expectation and differentiation. Using the previous result we have:

$$E_{t-1}\left[\frac{\partial^2 L\left(\hat{\sigma}_t^2, h_t^*\right)}{\partial h^2}\right] = E_{t-1}\left[c'\left(h_t^*\right)\left(\hat{\sigma}_t^2 - h_t^*\right) - c\left(h_t^*\right)\right] = -c\left(h_t^*\right)$$

which is non-negative iff $c\left(h_t^*\right)$ is non-positive. From assumption A4 we know that the optimum is in the interior of $\mathcal{H}$ and so we know that $c \neq 0$, and thus $c(h) < 0 \ \forall \ h \in \mathcal{H}$. To obtain the loss function corresponding to the given first derivative we simply integrate up:

$$\begin{aligned} L\left(\hat{\sigma}^2, h\right) &= \hat{\sigma}^2 \int c(h)\, dh - \int c(h)\, h dh \\ &= B\left(\hat{\sigma}^2\right) + \hat{\sigma}^2 C(h) - C(h)\, h + \int C(h)\, dh \\ &= \tilde{C}(h) + B\left(\hat{\sigma}^2\right) + C(h)\left(\hat{\sigma}^2 - h\right) \end{aligned}$$

where C is a strictly decreasing function (i.e. $C' \equiv c$ is negative) and $\tilde{C}$ is the anti-derivative of C. By assumption A3 both B and C are twice continuously differentiable. Thus $\mathcal{S}3 \Rightarrow \mathcal{S}1$.

Finally, we show that $\mathcal{S}2 \Rightarrow \mathcal{S}3$: by the definition of h_t^* we have

$$\begin{aligned} E_{t-1}\left[L\left(\hat{\sigma}_t^2, h_t^*\right)\right] &\leq E_{t-1}\left[L\left(\hat{\sigma}_t^2, \tilde{h}_t\right)\right] \text{ for any other } \tilde{h}_t \in \mathcal{F}_{t-1} \\ \text{so } E\left[L\left(\hat{\sigma}_t^2, h_t^*\right)\right] &\leq E\left[L\left(\hat{\sigma}_t^2, \tilde{h}_t\right)\right] \text{ by the LIE} \\ \text{and } E\left[L\left(\sigma_t^2, h_t^*\right)\right] &\leq E\left[L\left(\sigma_t^2, \tilde{h}_t\right)\right] \text{ since } L \text{ is robust under } \mathcal{S}2 \end{aligned}$$

But $L\left(\hat{\sigma}^2, h\right)$ has a unique minimum at $\hat{\sigma}^2 = h$, and if we set $\tilde{h}_t = \sigma_t^2 \in \mathcal{F}_{t-1}$ then it must be the case that $h_t^* = \sigma_t^2$. This completes the proof. ■

Proof of Proposition 3. Without loss of generality, we work below with loss functions that have been normalised to imply zero loss when the forecast error is zero: $L\left(\hat{\sigma}^2, h\right) = \tilde{C}(h) - \tilde{C}\left(\hat{\sigma}^2\right) + C(h)\left(\hat{\sigma}^2 - h\right)$.

(i) We want to find the general sub-set of loss functions that satisfy $L\left(\hat{\sigma}^2, h\right) = \tilde{L}\left(\hat{\sigma}^2 - h\right) \ \forall \left(\hat{\sigma}^2, h\right)$ for some function $\tilde{L}$. This condition implies

$$\begin{aligned} \frac{\partial L\left(\hat{\sigma}^2, h\right)}{\partial \hat{\sigma}^2} &= -\frac{\partial L\left(\hat{\sigma}^2, h\right)}{\partial h} \ \forall \left(\hat{\sigma}^2, h\right) \\ -C\left(\hat{\sigma}^2\right) + C(h) + C'(h)\left(\hat{\sigma}^2 - h\right) &= 0 \ \forall \left(\hat{\sigma}^2, h\right) \end{aligned}$$

Taking the derivative of both sides w.r.t. $\hat{\sigma}^2$ we obtain:

$$\begin{aligned} -C'\left(\hat{\sigma}^2\right) + C'(h) &= 0 \ \forall \left(\hat{\sigma}^2, h\right) \\ \text{which implies } C'(h) &= \kappa_1 \ \forall \ h \end{aligned}$$

and since we know C is strictly decreasing, we also have $\kappa_1 < 0$.

$$\begin{aligned} \text{so } C(h) &= \kappa_1 h + \kappa_2\left(\hat{\sigma}^2\right) \\ \tilde{C}(h) &= \frac{1}{2}\kappa_1 h^2 + \kappa_2\left(\hat{\sigma}^2\right) h + \kappa_3\left(\hat{\sigma}^2\right) \end{aligned}$$

where κ_2, κ_3 are constants of integration, and may be functions of $\hat{\sigma}^2$. Thus the loss function becomes

$$\begin{aligned} L\left(\hat{\sigma}^2, h\right) &= \frac{1}{2}\kappa_1 h^2 + \kappa_2\left(\hat{\sigma}^2\right) h + \kappa_3\left(\hat{\sigma}^2\right) \\ &\quad -\frac{1}{2}\kappa_1 \hat{\sigma}^4 - \kappa_2\left(\hat{\sigma}^2\right)\hat{\sigma}^2 - \kappa_3\left(\hat{\sigma}^2\right) \\ &\quad + \left(\kappa_1 h + \kappa_2\left(\hat{\sigma}^2\right)\right)\left(\hat{\sigma}^2 - h\right) \\ &= -\frac{1}{2}\kappa_1\left(\hat{\sigma}^2 - h\right)^2 \end{aligned}$$

Since proportionality constants do not affect the loss function, we find that the only loss function that depends on $\left(\hat{\sigma}^2, h\right)$ only through the forecast error, $\hat{\sigma}^2 - h$, is the MSE loss function.

(ii) We next want to find the general sub-set of loss functions that satisfy $L\left(\hat{\sigma}^2, h\right) = \tilde{L}\left(\hat{\sigma}^2/h\right)\ \forall \left(\hat{\sigma}^2, h\right)$ for some function $\tilde{L}$. Note that this condition implies that L is homogeneous of degree zero. Using Proposition 4 below, this implies that the loss function must be of the form:

$$L\left(\hat{\sigma}^2, h\right) = \frac{\hat{\sigma}^2}{h} - \log\frac{\hat{\sigma}^2}{h} - 1$$

which is the QLIKE loss function up to additive and multiplicative constants. ■

Proof of Proposition 4. (i) It is obvious $L(h, h; b) = 0\ \forall\ h \in \mathcal{H}$. We now show that all three of these loss functions are of the form in Proposition 2.

$b \notin \{-1, -2\}$: $C(h) = -(b+1)^{-1} h^{b+1}$, $\tilde{C}(h) = -(b+1)^{-1}(b+2)^{-1} h^{b+2}$, $B\left(\hat{\sigma}^2\right) = (b+1)^{-1}(b+2)^{-1}\hat{\sigma}^{2b+4}$.

$b = -1$: $C(h) = -\log h$, $\tilde{C}(h) = h - h\log h$, $B\left(\hat{\sigma}^2\right) = \hat{\sigma}^2\log\hat{\sigma}^2 - \hat{\sigma}^2$.

$b = -2$: $C(h) = h^{-1}$, $\tilde{C}(h) = \log h$, $B\left(\hat{\sigma}^2\right) = -\log\hat{\sigma}^2$.

(ii) We seek the subset of robust loss functions that are homogeneous of order k : $L\left(a\hat{\sigma}^2, ah\right) = a^k L\left(\hat{\sigma}^2, h\right)\ \forall\ a > 0$. Let

$$\begin{aligned} \lambda\left(\hat{\sigma}^2, h\right) &\equiv \partial L\left(\hat{\sigma}^2, h\right)/\partial h \\ &= C'(h)\left(\hat{\sigma}^2 - h\right) \quad \text{for robust loss functions.} \end{aligned}$$

Since L is homogeneous of order k, λ is homogeneous of order $(k-1)$. This implies $\lambda\left(a\hat{\sigma}^2, ah\right) = a^{k-1}\lambda\left(\hat{\sigma}^2, h\right) = a^{k-1}C'(h)\left(\hat{\sigma}^2 - h\right)$, while direct substitution yields $\lambda\left(a\hat{\sigma}^2, ah\right) = aC'(ah)\left(\hat{\sigma}^2 - h\right)$. Thus $C'(ah) = a^{k-2}C'(h)\ \forall\ a > 0$, that is, C' is homogeneous of order $(k-2)$.

Next we apply Euler's theorem to C': $C''(h)\,h = (k-2)\,C'(h)\ \ \forall h > 0\,,$ and so

$$(2-k)\,C'(h) + C''(h)\,h = 0$$

We can solve this first-order differential equation to find:

$$C'(h) = \gamma h^{k-2}$$

where γ is an unknown scalar. Since $C' < 0$ we know that $\gamma < 0$, and as this is just a scaling parameter we set it to -1 without loss of generality.

$$\begin{aligned} C'(h) &= -h^{k-2} \\ C(h) &= \begin{cases} \frac{1}{1-k}h^{k-1} + z_1 & k \neq 1 \\ -\log h + z_1 & k = 1 \end{cases} \\ \tilde{C}(h) &= \begin{cases} z_1 h + \frac{1}{k(1-k)}h^k + z_2 & k \notin \{0,1\} \\ z_1 h + h - h\log h + z_2 & k = 1 \\ z_1 h + \log h + z_2 & k = 0 \end{cases} \end{aligned}$$

where z_1 and z_2 are constants of integration. Finally, we substitute the expressions for C and $\tilde{C}$ into equation (25) and simplify to obtain the loss functions in equation (27) with $k = b + 2$. ■

Proof of Proposition 5. (i) If L is homogeneous then $L\left(a\hat{\sigma}^2, ah\right) = a^k L\left(\hat{\sigma}^2, h\right)\ \ \forall a > 0$ for some k. Then $E\left[L\left(a\hat{\sigma}_t^2, ah_{1t}\right)\right] \geq E\left[L\left(a\hat{\sigma}_t^2, ah_{2t}\right)\right] \Leftrightarrow E\left[a^k L\left(\hat{\sigma}_t^2, h_{1t}\right)\right] \geq E\left[a^k L\left(\hat{\sigma}_t^2, h_{2t}\right)\right] \Leftrightarrow E\left[L\left(\hat{\sigma}_t^2, h_{1t}\right)\right] \geq E\left[L\left(\hat{\sigma}_t^2, h_{2t}\right)\right]$, for any $a > 0$.

(ii) Here we need only provide an example. Consider the following stylised case: $\sigma_t^2 = 1\ \forall t$, $(h_{1t}, h_{2t}) = (\gamma_1, \gamma_2)\ \forall t,$ and $\hat{\sigma}_t^2$ is such that $E_{t-1}\left[\hat{\sigma}_t^2\right] = 1$ a.s. $\forall t$. As a robust but non-homogeneous loss we will use the one generated by the following specification for C':

$$\begin{aligned} C'(h) &= -\log(1+h) \\ \text{so}\ \ C(h) &= h - (1+h)\log(1+h) \\ \text{and}\ \ \tilde{C}(h) &= \frac{1}{4}\left[h(3h+2) - 2(1+h)^2\log(1+h)\right] \end{aligned}$$

For small h this loss function resembles the the $b = 1$ loss function from Proposition 4 (up to a scaling constant), but for medium to large h this loss function does not correspond to any in Proposition 4.

Given this set-up, we have

$$\begin{aligned} E\left[L\left(a\hat{\sigma}_t^2, ah_{it}\right)\right] &= \frac{1}{4}\left[a\gamma_i\left(3a\gamma_i+2\right)-2\left(1+a\gamma_i\right)^2\log\left(1+a\gamma_i\right)\right]-E\left[\tilde{C}\left(a\hat{\sigma}_t^2\right)\right] \\ &\quad +a\left[a\gamma_i-\left(1+a\gamma_i\right)\log\left(1+a\gamma_i\right)\right]\left(1-\gamma_i\right) \end{aligned}$$

Then define

$$\begin{aligned} d_t\left(\gamma_1,\gamma_2,a\right) &\equiv L\left(a\hat{\sigma}_t^2, a\gamma_1\right)-L\left(a\hat{\sigma}_t^2, a\gamma_2\right) \\ E\left[d_t\left(\gamma_1,\gamma_2,a\right)\right] &= \frac{a}{4}\left(\gamma_1-\gamma_2\right)\left(2-4a-a\left(\gamma_1+\gamma_2\right)\right) \\ &\quad +\frac{1}{2}\left(a^2\left(\gamma_1-1\right)^2-\left(1+a\right)^2\right)\log\left(1+a\gamma_1\right) \\ &\quad -\frac{1}{2}\left(a^2\left(\gamma_2-1\right)^2-\left(1+a\right)^2\right)\log\left(1+a\gamma_2\right) \end{aligned}$$

Then note that $E\left[d_t\left(1/3,3/2,1\right)\right]=-0.0087$, and so the first forecast has lower expected loss than the second using the "original" scaling of the data. But $E\left[d_t\left(1/3,3/2,2\right)\right]=0.0061$, and so if all variables are multiplied by 2 then the second forecast has lower expected loss than the first. ■

Proof of Proposition 6. (i) Follows directly from Exercise 5.21 of White (1999).

(ii) For $b\notin\{-1,-2\}$ we have:

$$\begin{aligned} d_t\left(b\right) &= \frac{1}{b+2}\left(h_{1t}^{b+2}-h_{2t}^{b+2}\right)-\frac{1}{b+1}\hat{\sigma}_t^2\left(h_{1t}^{b+1}-h_{2t}^{b+1}\right) \\ \text{so } d_t\left(b\right)^2 &= \frac{1}{\left(b+2\right)^2}\left(h_{1t}^{2b+4}+h_{2t}^{2b+4}-2h_{1t}^{b+2}h_{2t}^{b+2}\right) \\ &\quad +\frac{1}{\left(b+1\right)^2}\hat{\sigma}_t^4\left(h_{1t}^{2b+2}+h_{2t}^{2b+2}-2h_{1t}^{b+1}h_{2t}^{b+1}\right) \\ &\quad -\frac{2}{\left(b+1\right)\left(b+2\right)}\hat{\sigma}_t^2\left(h_{1t}^{2b+3}+h_{2t}^{2b+3}-h_{1t}^{b+2}h_{2t}^{b+1}-h_{1t}^{b+1}h_{2t}^{b+2}\right) \end{aligned}$$

The largest terms in this expression are: (a) h_{it}^{2b+4}, (b) $\hat{\sigma}_t^4h_{it}^{2b+2}$ and (c) $\hat{\sigma}_t^2h_{it}^{2b+3}$ for $i=1,2$. The expectation of first term is finite by assumption. (b) If $b>-1$, then by Hölder's inequality:

$$\begin{aligned} E\left[\hat{\sigma}_t^4h_{it}^{2b+2}\right] &\leq E\left[\left(\hat{\sigma}_t^4\right)^{b+2}\right]^{1/(b+2)}E\left[\left(h_{it}^{2b+2}\right)^{(b+2)/(b+1)}\right]^{(b+1)/(b+2)} \\ &= E\left[\hat{\sigma}_t^{4b+8}\right]^{1/(b+2)}E\left[h_{it}^{2b+4}\right]^{(b+1)/(b+2)} \\ &< \infty \text{ by assumption.} \end{aligned}$$

If $b<-1$, then we will make use of the assumption that $\inf_t h_{it}\equiv c_i>0$ for $i=1,2$. This assumption implies that h_{it}^{-1} is bounded below by zero and above by c_i^{-1}, and thus all moments of

h_{it}^{-1} exist.

$$\begin{aligned} E\left[\hat{\sigma}_t^4 h_{it}^{2b+2}\right] &\leq E\left[\left(\hat{\sigma}_t^4\right)^{(4+\delta)/4}\right]^{4/(4+\delta)} E\left[\left(h_{it}^{2b+2}\right)^{(4+\delta)/\delta}\right]^{\delta/(4+\delta)}, \text{ for } \delta > 0 \\ &= E\left[\hat{\sigma}_t^{4+\delta}\right]^{4/(4+\delta)} E\left[h_{it}^{2(b+1)(4+\delta)/\delta}\right]^{\delta/(4+\delta)} \end{aligned}$$

which is finite as $E\left[\hat{\sigma}_t^{4+\delta}\right] < \infty$ by assumption, and $E\left[h_{it}^{-M}\right] < \infty$ for all $0 \leq M < \infty$ since h_{it}^{-1} is a bounded random variable. (c) If $b > -1.5$, then

$$\begin{aligned} E\left[\hat{\sigma}_t^2 h_{it}^{2b+3}\right] &\leq E\left[\hat{\sigma}_t^{4b+8}\right]^{1/(2b+4)} E\left[\left(h_{it}^{2b+3}\right)^{(2b+4)/(2b+3)}\right]^{(2b+3)/(2b+4)} \\ &= E\left[\hat{\sigma}_t^{4b+8}\right]^{1/(2b+4)} E\left[h_{it}^{2b+4}\right]^{(2b+3)/(2b+4)} \\ &< \infty \text{ by assumption.} \end{aligned}$$

If $b < -1.5$,

$$\begin{aligned} E\left[\hat{\sigma}_t^2 h_{it}^{2b+3}\right] &\leq E\left[\left(\hat{\sigma}_t^2\right)^{(4+\delta)/2}\right]^{2/(4+\delta)} E\left[\left(h_{it}^{2b+3}\right)^{(4+\delta)/(2+\delta)}\right]^{(2+\delta)/(4+\delta)} \\ &= E\left[\hat{\sigma}_t^{4+\delta}\right]^{2/(4+\delta)} E\left[h_{it}^{(2b+3)(4+\delta)/(2+\delta)}\right]^{(2+\delta)/(4+\delta)} \end{aligned}$$

which is finite as $E\left[\hat{\sigma}_t^{4+\delta}\right] < \infty$ by assumption, and $E\left[h_{it}^{-M}\right] < \infty$ for all $0 \leq M < \infty$ since h_{it}^{-1} is a bounded random variable.

Now consider $b = -1$. Here we have

$$\begin{aligned} d_t &= h_{1t} - h_{2t} - \hat{\sigma}_t^2 \log \frac{h_{1t}}{h_{2t}} \\ \text{so } d_t^2 &= (h_{1t} - h_{2t})^2 + \hat{\sigma}_t^4 \left(\log(h_{1t}) - \log(h_{2t})\right)^2 \\ &\quad -2\hat{\sigma}_t^2 \left(\log(h_{1t}) - \log(h_{2t})\right)(h_{1t} - h_{2t}) \end{aligned}$$

The largest terms in this expression are: (a) h_{it}^2, (b) $\hat{\sigma}_t^4 (\log h_{it})^2$ and (c) $\hat{\sigma}_t^2 h_{it} \log h_{it}$. The first term is finite by assumption. For (b) we have

$$\begin{aligned} E\left[\hat{\sigma}_t^4 (\log h_{it})^2\right] &\leq E\left[\left(\hat{\sigma}_t^4\right)^{(e+1)/e}\right]^{e/(e+1)} E\left[(\log h_{it})^{2(e+1)}\right]^{1/(e+1)} \\ &= E\left[\hat{\sigma}_t^{4(e+1)/e}\right]^{e/(e+1)} E\left[(\log h_{it})^{2(e+1)}\right]^{1/(e+1)} \end{aligned}$$

The first term on the right-hand side is finite by assumption. For the second term note:

$$\begin{aligned}
(\log h)^k &\leq h^{k/e} \;\forall\; h \geq 1 \text{ for } k > 0 \\
\text{And } \; E\left[(\log h_{it})^k\right] &\equiv \int_{c_i}^{\infty} (\log h_{it})^k f_{it}(h_{it})\, dh_{it} \\
&= \int_{c_i}^{1} (\log h_{it})^k f_{it}(h_{it})\, dh_{it} + \int_{1}^{\infty} (\log h_{it})^k f_{it}(h_{it})\, dh_{it} \\
\left|\int_{c_i}^{1} (\log h_{it})^k f_{it}(h_{it})\, dh_{it}\right| &< \infty \text{ for } k > 0 \text{ since } c_i > 0
\end{aligned}$$

$$\begin{aligned}
\int_{1}^{\infty} (\log h_{it})^k f_{it}(h_{it})\, dh_{it} &\leq \int_{1}^{\infty} h_{it}^{k/e} f_{it}(h_{it})\, dh_{it} \\
&\leq \int_{c_i}^{\infty} h_{it}^{k/e} f_{it}(h_{it})\, dh_{it} \\
&\equiv E\left[h_{it}^{k/e}\right]
\end{aligned}$$

So $E\left[(\log h_{it})^{2(e+1)}\right] < \infty$ if $E\left[h_{it}^{2(e+1)/e}\right] < \infty$, which holds by assumption. For (c) we have

$$\begin{aligned}
E\left[\hat{\sigma}_t^2 h_{it} \log h_{it}\right] &\leq E\left[\left(\hat{\sigma}_t^2\right)^{(2e+1)/e}\right]^{e/(2e+1)} E\left[h_{it}^{(2e+1)/e}\right]^{e/(2e+1)} E\left[(\log h_{it})^{2e+1}\right]^{1/(2e+1)} \\
&< \infty
\end{aligned}$$

As the first two terms on the right-hand side are finite by assumption, and the final term is finite given that the second term is finite and $c_i > 0$.

Finally consider $b = -2$. We have

$$\begin{aligned}
d_t &= \hat{\sigma}_t^2 \left(\frac{1}{h_{1t}} - \frac{1}{h_{2t}}\right) + \log \frac{h_{1t}}{h_{2t}} \\
\text{so } \; d_t^2 &= \hat{\sigma}_t^4 \left(\frac{1}{h_{1t}} - \frac{1}{h_{2t}}\right)^2 + 2\left(\log h_{1t} - \log h_{2t}\right)^2 \\
&\quad + 2\hat{\sigma}_t^2 \left(\frac{1}{h_{1t}} - \frac{1}{h_{2t}}\right) \log h_{1t} - 2\hat{\sigma}_t^2 \left(\frac{1}{h_{1t}} - \frac{1}{h_{2t}}\right) \log h_{2t}
\end{aligned}$$

with largest terms: (a) $\hat{\sigma}_t^4 h_{it}^{-2}$, (b) $(\log h_{it})^2$, (c) $\hat{\sigma}_t^2 h_{it}^{-1} \log h_{it}$, and (d) $\hat{\sigma}_t^2 h_{jt}^{-1} \log h_{it}$ for $i = 1, 2$ and $j \neq i$. For term (a) we have

$$\begin{aligned}
E\left[\hat{\sigma}_t^4 \frac{1}{h_{it}^2}\right] &\leq E\left[\left(\hat{\sigma}_t^4\right)^{(4+\delta)/4}\right]^{4/(4+\delta)} E\left[\left(\frac{1}{h_{it}^2}\right)^{(4+\delta)/\delta}\right]^{\delta/(4+\delta)}, \text{ for } \delta > 0 \\
&= E\left[\hat{\sigma}_t^{4+\delta}\right]^{4/(4+\delta)} E\left[h_{it}^{-2(4+\delta)/\delta}\right]^{\delta/(4+\delta)}
\end{aligned}$$

The first term on the right-hand side is finite by assumption and the second term is finite since $c_i > 0$. For term (b) recall from above that if $c_i > 0$ then $E\left[(\log h_{it})^k\right] < \infty$ if $E\left[h_{it}^{k/e}\right] < \infty$ for

$k > 0$. So $E\left[(\log h_{it})^2\right] < \infty$ is finite if $E\left[h_{it}^{2/e}\right] < \infty$, which holds by assumption. For term (c) we use

$$\begin{aligned} E\left[\hat{\sigma}_t^2 h_{it}^{-1} \log h_{it}\right] &\leq E\left[\left(\hat{\sigma}_t^2\right)^{(4+\delta)/2}\right]^{2/(4+\delta)} E\left[\left(h_{it}^{-1} \log h_{it}\right)^{(4+\delta)/(2+\delta)}\right]^{(2+\delta)/(4+\delta)} \text{ for } \delta > 0 \\ &= E\left[\hat{\sigma}_t^{4+\delta}\right]^{2/(4+\delta)} E\left[\left(h_{it}^{-1} \log h_{it}\right)^{(4+\delta)/(2+\delta)}\right]^{(2+\delta)/(4+\delta)} \end{aligned}$$

which is finite as the first term is finite by assumption and the second term is finite since $h_{it}^{-1} \log h_{it}$ is a bounded random variable if $c_i > 0$. For the term in (d) we have

$$\begin{aligned} E\left[\hat{\sigma}_t^2 h_{jt}^{-1} \log h_{it}\right] &\leq E\left[\left(\hat{\sigma}_t^2 \log h_{it}\right)^{(1+\delta/4)}\right]^{1/(1+\delta/4)} E\left[h_{jt}^{-(4+\delta)/\delta}\right]^{\delta/(4+\delta)}, \text{ for } \delta > 0 \\ &\leq \left(E\left[\left(\hat{\sigma}_t^{2(1+\delta/4)}\right)^2\right]^{1/2} E\left[\left((\log h_{it})^{(1+\delta/4)}\right)^2\right]^{1/2}\right)^{1/(1+\delta/4)} E\left[h_{jt}^{-(4+\delta)/\delta}\right]^{\delta/(4+\delta)} \\ &= \left(E\left[\hat{\sigma}_t^{4+\delta}\right]^{1/2} E\left[(\log h_{it})^{2+\delta/2}\right]^{1/2}\right)^{1/(1+\delta/4)} E\left[h_{jt}^{-(4+\delta)/\delta}\right]^{\delta/(4+\delta)} \end{aligned}$$

The first term is finite by assumption, and the third term is finite as $c_i > 0$. The second term is finite if $E\left[h_{it}^{(2+\delta/2)/e}\right]$ is finite, and $E\left[h_{it}^{(2+\delta/2)/e}\right] < E\left[h_{it}^{2/e+\delta}\right]$, which is finite by assumption. This completes the proof. ■

8 Tables and Figures

Table 1: Optimal forecasts under various loss functions

Loss function	Optimal forecast, h_t^*				
		$r_t\|\mathcal{F}_{t-1} \sim Student's\ t\left(0,\sigma_t^2,\nu\right)$			
	$r_t\|\mathcal{F}_{t-1} \sim F_t\left(0,\sigma_t^2\right)$	ν	$\nu=6$	$\nu=10$	$\nu\to\infty$
MSE	σ_t^2	σ_t^2	σ_t^2	σ_t^2	σ_t^2
QLIKE	σ_t^2	σ_t^2	σ_t^2	σ_t^2	σ_t^2
MSE-LOG	$\exp\left\{E_{t-1}\left[\log\varepsilon_t^2\right]\right\}\sigma_t^2$	$\exp\left\{\Psi\left(\frac{1}{2}\right)-\Psi\left(\frac{\nu}{2}\right)\right\}(\nu-2)\,\sigma_t^2$	$0.22\sigma_t^2$	$0.25\sigma_t^2$	$0.28\sigma_t^2$
MSE-SD	$\left(E_{t-1}\left[\left\|\varepsilon_t\right\|\right]\right)^2\sigma_t^2$	$\frac{\nu-2}{\pi}\left(\Gamma\left(\frac{\nu-1}{2}\right)/\Gamma\left(\frac{\nu}{2}\right)\right)^2\sigma_t^2$	$0.56\sigma_t^2$	$0.60\sigma_t^2$	$0.64\sigma_t^2$
MSE-prop	$Kurtosis_{t-1}\left[r_t\right]\sigma_t^2$	$3\frac{\nu-2}{\nu-4}\sigma_t^2$	$6.00\sigma_t^2$	$4.00\sigma_t^2$	$3.00\sigma_t^2$
MAE	$Median_{t-1}\left[r_t^2\right]$	$\frac{\nu-2}{\nu}Median\left[F_{1,\nu}\right]\sigma_t^2$	$0.34\sigma_t^2$	$0.39\sigma_t^2$	$0.45\sigma_t^2$
MAE-LOG	$Median_{t-1}\left[r_t^2\right]$	$\frac{\nu-2}{\nu}Median\left[F_{1,\nu}\right]\sigma_t^2$	$0.34\sigma_t^2$	$0.39\sigma_t^2$	$0.45\sigma_t^2$
MAE-SD	$Median_{t-1}\left[r_t^2\right]$	$\frac{\nu-2}{\nu}Median\left[F_{1,\nu}\right]\sigma_t^2$	$0.34\sigma_t^2$	$0.39\sigma_t^2$	$0.45\sigma_t^2$
MAE-prop†	n/a	$\left(2.36+\frac{1.00}{\nu}+\frac{7.78}{\nu^2}\right)\sigma_t^2$	$2.73\sigma_t^2$	$2.55\sigma_t^2$	$2.36\sigma_t^2$

Notes: This table presents the forecast that minimises the conditional expected loss when the squared return is used as a volatility proxy. That is, h_t^* minimises $E_{t-1}\left[L\left(r_t^2,h\right)\right]$, for various loss functions L. The first column presents the solutions when returns have an arbitrary conditional distribution F_t with mean zero and conditional variance σ_t^2, the second, third, and fourth columns present results with returns have the standardised Student's t distribution, and the final column presents the solutions when returns are conditionally normally distributed. Γ is the gamma function and Ψ is the digamma function. †The expressions given for MAE-prop are based on a numerical approximation, see Appendix 1 for details.

Table 2: Optimal forecasts under various loss functions, using realised volatility and range

Loss function	**Volatility proxy**					
		Realised volatility				
	Range	Arbitrary m	$m=1$	$m=13$	$m=78$	$m\to\infty$
MSE	σ_t^2	σ_t^2	σ_t^2	σ_t^2	σ_t^2	σ_t^2
QLIKE	σ_t^2	σ_t^2	σ_t^2	σ_t^2	σ_t^2	σ_t^2
MSE-LOG†	$0.85\sigma_t^2$	$e^{-1.2741/m}\sigma_t^2$	$0.28\sigma_t^2$	$0.91\sigma_t^2$	$0.98\sigma_t^2$	σ_t^2
MSE-SD	$0.92\sigma_t^2$	$\frac{1}{m}\left(E\left[\sqrt{\chi_m^2}\right]\right)^2\sigma_t^2$	$0.56\sigma_t^2$	$0.96\sigma_t^2$	$0.99\sigma_t^2$	σ_t^2
MSE-prop	$1.41\sigma_t^2$	$\left(1+\frac{2}{m}\right)\sigma_t^2$	$3.00\sigma_t^2$	$1.15\sigma_t^2$	$1.03\sigma_t^2$	σ_t^2
MAE	$0.83\sigma_t^2$	$\frac{1}{m}Median\left[\chi_m^2\right]\sigma_t^2$	$0.45\sigma_t^2$	$0.95\sigma_t^2$	$0.99\sigma_t^2$	σ_t^2
MAE-LOG	$0.83\sigma_t^2$	$\frac{1}{m}Median\left[\chi_m^2\right]\sigma_t^2$	$0.45\sigma_t^2$	$0.95\sigma_t^2$	$0.99\sigma_t^2$	σ_t^2
MAE-SD	$0.83\sigma_t^2$	$\frac{1}{m}Median\left[\chi_m^2\right]\sigma_t^2$	$0.45\sigma_t^2$	$0.95\sigma_t^2$	$0.99\sigma_t^2$	σ_t^2
MAE-prop†	$1.19\sigma_t^2$	$\left(1+\frac{1.3624}{m}\right)\sigma_t^2$	$2.36\sigma_t^2$	$1.10\sigma_t^2$	$1.02\sigma_t^2$	σ_t^2

Notes: This table presents the forecast that minimises the conditional expected loss when the range or realised volatility is used as a volatility proxy. That is, h_t^* minimises $E_{t-1}\left[L\left(\hat{\sigma}_t^2,h\right)\right]$, for $\hat{\sigma}_t^2=RG_t^{*2}$ or $\hat{\sigma}_t^2=RV_t$, for various loss functions L. In all cases returns are assumed to be generated as a zero mean Brownian motion with constant volatility within each trade day and no jumps. The cases of $m=1,\ 13,\ 78$ correspond to the use of daily squared returns, realised variance with 30-minute returns and realised variance with 5-minute returns respectively. The case that $m\to\infty$ corresponds to the case where the conditional variance is observable ex-post without error. †For the MSE-LOG and MAE-prop loss functions we used simulations, numerical integration and numerical optimisation to obtain the expressions given. Details on the computation of the figures in this table are given in Appendix 1.

Table 3: Mincer-Zarnowitz tests of the volatility forecasts

		Volatility proxy			
Volatility Model		*Daily squared return*	*65-min realised vol*	*15-min realised vol*	*5-min realised vol*
Rolling	$\hat{\beta}_0$	2.13*	1.82*	2.17*	2.33*
	$(s.e.)$	(0.48)	(0.36)	(0.40)	(0.40)
window	$\hat{\beta}_1$	0.55*	0.55*	0.50*	0.53*
	$(s.e.)$	(0.09)	(0.07)	(0.07)	(0.07)
	χ^2_2-stat	25.63*	44.32*	53.78*	43.86*
	pval	0.00	0.00	0.00	0.00
	$\hat{\beta}_0$	2.39*	2.15*	2.20*	2.43*
	$(s.e.)$	(0.46)	(0.40)	(0.44)	(0.42)
RiskMetrics	$\hat{\beta}_1$	0.50*	0.48*	0.50*	0.51*
	$(s.e.)$	(0.09)	(0.08)	(0.09)	(0.09)
	χ^2_2-stat	32.99*	38.51*	30.24*	35.93*
	pval	0.00	0.00	0.00	0.00

Notes: This table presents the results of Mincer-Zarnowitz (MZ) tests of two IBM equity volatility forecasts: a 60-day rolling window forecast, and a RiskMetrics forecast. The sample period is January 1994 to December 2003. The null hypothesis in the MZ test is that $\beta_0 = 0$ and $\beta_1 = 1$. We present the parameter estimates and Newey-West standard errors, and mark any parameter estimates that are significantly different from their hypothesised values at the 0.05 level with an asterisk. We also present the results of a χ^2_2 test of the joint parameter restriction and the p-value associated with the joint test statistic. A p-value of less than 0.05 indicates a rejection of the null, and thus evidence against the optimality of the volatility forecast. These statistics are marked with an asterisk.

Table 4: Comparison of rolling window and RiskMetrics forecasts

	Volatility proxy			
Loss function	*Daily squared return*	*65-min realised vol*	*15-min realised vol*	*5-min realised vol*
b = 1	-1.58	-1.66	-1.30	-1.35
b = 0 (MSE)	-0.59	-0.80	-0.03	-0.13
b = -1	1.30	1.04	1.65	-1.55
b = -2 (QLIKE)	1.94	2.21*	2.73*	2.41*
b = -5	-0.17	0.25	1.63	0.65

Notes: This table presents the t-statistics from Diebold-Mariano-West tests of equal predictive accuracy for a 60-day rolling window forecast and a RiskMetrics forecast, for IBM over the period January 1994 to December 2003. A t-statistic greater than 1.96 in absolute value indicates a rejection of the null of equal predictive accuracy at the 0.05 level. These statistics are marked with an asterisk. The sign of the t-statistics indicates which forecast performed better for each loss function: a positive t-statistic indicates that the rolling window forecast produced larger average loss than the RiskMetrics forecast, while a negative sign indicates the opposite.

References

[1] Alizadeh, Sassan, Brandt, Michael W., and Diebold, Francis X., 2002, Range-Based Estimation of Stochastic Volatility Models, *Journal of Finance*, 57(3), 1047-1091.

[2] Andersen, Torben G., Benzoni, Luca, and Lund, Jesper, 2002, An Empirical Investigation of Continuous-Time Equity Return Models, *Journal of Finance*, 57(3), 1239-1284.

[3] Andersen, Torben G., and Bollerslev, Tim, 1998, Answering the Skeptics: Yes, Standard Volatility Models Do Provide Accurate Forecasts, *International Economic Review*, 39, 885-905.

[4] Andersen, Torben G., Bollerslev, Tim, and Lange, Steve, 1999, Forecasting Financial Market Volatility: Sample Frequency Vis-à-vis Forecast Horizon, *Journal of Empirical Finance*, 6, 457-477.

[5] Andersen, Torben, Bollerslev, Tim, and Diebold, Francis X., 2002, "Parametric and Nonparametric Volatility Measurement," forthcoming in L.P. Hansen and Y. Ait-Sahalia (eds.), *Handbook of Financial Econometrics*, Amsterdam: North-Holland.

[6] Andersen, Torben, Bollerslev, Tim, Diebold, Francis X., and Ebens, Heiko, 2001a, The Distribution of Realized Stock Return Volatility, *Journal of Financial Economics*, 61, 43-76.

[7] Andersen, Torben G., Bollerslev, Tim, Diebold, Francis X. and Labys, Paul, 2001b, The Distribution of Realized Exchange Rate Volatility, *Journal of the American Statistical Association*, 96, 42-55.

[8] Andersen, Torben G., Bollerslev, Tim, Diebold, Francis X. and Labys, Paul, 2003, Modeling and Forecasting Realized Volatility, *Econometrica*, 71(2), 579-625.

[9] Andersen, Torben G., Bollerslev, Tim, and Meddahi, Nour, 2004, Analytic Evaluation of Volatility Forecasts, *International Economic Review*, 45, 1079-1110.

[10] Andersen, Torben G., Bollerslev, Tim, and Meddahi, Nour, 2005a, Correcting the Errors: Volatility Forecast Evaluation Using High-Frequency Data and Realized Volatilities, *Econometrica*, 73(1), 279-296.

[11] Andersen, Torben G., Bollerslev, Tim, Christoffersen, Peter F., and Diebold, Francis X., 2005b, Volatility and Correlation Forecasting, in the *Handbook of Economic Forecasting*, G. Elliott, C.W.J. Granger and A. Timmermann ed.s, North Holland Press, Amsterdam.

[12] Ball, Clifford A., and Torous, Walter N., 1984, The Maximum Likelihood Estimation of Security Price Volatility: Theory, Evidence, and Application to Option Pricing, *Journal of Business*, 57(1), 97-112.

[13] Barndorff-Nielsen, Ole E., and Shephard, Neil, 2002, Econometric Analysis of Realised Volatility and Its Use in Estimating Stochastic Volatility Models, *Journal of the Royal Statistical Society*, Series B, 64, 253-280.

[14] Barndorff-Nielsen, Ole E., and Shephard, Neil, 2004, Econometric Analysis of Realized Covariation: High Frequency Based Covariance, Regression and Correlation in Financial Economics, *Econometrica*, 72(3), 885-925.

[15] Bollerslev, Tim, and Ghysels, Eric, 1994, Periodic Autoregressive Conditional Heteroscedasticity, *Journal of Business and Economic Statistics*, 14(2), 139-151.

[16] Bollerslev, Tim, and Wright, Jonathan H., 2001, High-Frequency Data, Frequency Domain Inference, and Volatility Forecasting, *Review of Economics and Statistics*, 83(5), 596-602.

[17] Chen, Runquan and Patton, Andrew J., 2006, Volatility Forecast Evaluation and Comparison, work-in-progress, To be published in T.G. Andersen, R.A. Davis, J.-P. Kreiss and T. Mikosch (eds.) *Handbook of Financial Time Series*, Springer Verlag.

[18] Christensen, Kim, and Podolskij, Mark, 2005, Asymptotic Theory for Range-Based Estimation of Integrated Variance of a Continuous Semi-Martingale, working paper, Aarhus School of Business.

[19] Christoffersen, Peter and Diebold, Francis X., 1997, Optimal prediction under asymmetric loss, *Econometric Theory,* 13, 808-817.

[20] Christoffersen, Peter, and Jacobs, Kris, 2004, The Importance of the Loss Function in Option Valuation, *Journal of Financial Economics*, 72, 291-318.

[21] Clements, Michael P., 2005, *Evaluating Econometric Forecasts of Economic and Financial Variables*, Palgrave MacMillan, United Kingdom.

[22] Cowles, Alfred, 1933, Can Stock Market Forecasters Forecast?, *Econometrica*, 1(3), 309-324.

[23] Diebold, Francis X., and Mariano, Roberto S., 1995, Comparing Predictive Accuracy, *Journal of Business and Economic Statistics,* 13(3), 253-263.

[24] Diebold, Francis X., and Lopez, Jose A., 1996, Forecast Evaluation and Combination, in G.S. Maddala and C.R. Rao (eds.), *Handbook of Statistics*, Amsterdam: North-Holland, 241-268.

[25] Engle, Robert F., 1982, Autoregressive Conditional Heteroskedasticity With Estimates of the Variance of U.K. Inflation, *Econometrica,* 50, 987-1008.

[26] Engle, Robert F., 1993, A Comment on Hendry and Clements on the Limitations of Comparing Mean Square Forecast Errors, *Journal of Forecasting*, 12, 642-644.

[27] Engle, R.F., C.-H. Hong, A. Kane and J. Noh, 1993, Arbitrage valuation of variance forecasts with simulated options, in: D. Chance and R. Tripp, eds., *Advances in Futures and Options Research*, JIA Press, Greenwich, U.S.A.

[28] Feller, W., 1951, The Asymptotic Distribution of the Range of Sums of Random Variables, *Annals of Mathematical Statistics*, 22, 427-432.

[29] Garman, Mark B., and Klass, Michael J., 1980, On the Estimation of Security Price Volatilities from Historical Data, *Journal of Business*, 53(1), 67-78.

[30] Gonçalves, Sílvia, and Meddahi, Nour, 2005, Bootstrapping Realized Volatility, working paper, Université de Montréal.

[31] Gourieroux, C., Monfort, A., and Trognon, A., 1984, Pseudo Maximum Likelihood Methods: Theory, *Econometrica*, 52(3), 681-700.

[32] Gourieroux, C., Monfort, A. and Renault, E., 1987, Consistent M-Estimators in a Semi-Parametric Model, CEPREMAP working paper 8720.

[33] Gourieroux, C. and Monfort, A., 1996, *Statistics and Econometric Models, Volume 1*, translated from the French by Q. Vuong, Cambridge University Press, Great Britain.

[34] Granger, C.W.J., 1969, Prediction with a generalized cost function, *OR*, 20, 199-207.

[35] Hamilton, James D., and Susmel, Rauli, 1994, Autoregressive Conditional Heteroskedasticity and Changes in Regime, *Journal of Econometrics*, 64(1-2), 307-333.

[36] Hansen, Peter Reinhard, and Lunde, Asger, 2005, A Forecast Comparison of Volatility Models: Does Anything Beat a GARCH(1,1)?, *Journal of Applied Econometrics*, 20(7), 873-889.

[37] Hansen, Peter R., and Lunde, Asger, 2006, Consistent Ranking of Volatility Models, *Journal of Econometrics*, 131(1-2), 97-121.

[38] Harvey, Andrew, Ruiz, Esther, and Shephard, Neil, 1994, Multivariate Stochastic Volatility Models, *Review of Economic Studies*, 61, 247-264.

[39] Huber, P.J., 1981, *Robust Statistics*, Wiley, New York, U.S.A.

[40] Komunjer, I., and Vuong, Q., 2004, Efficient Conditional Quantile Estimation, working paper.

[41] Lamoureux, C.G. and Lastrapes, W.D., 1993, Forecasting Stock Return Variance: Toward an Understanding of Stochastic Implied Volatilities, *Review of Financial Studies*, 6(2), 293-326.

[42] Martens, Martin, and van Dijk, Dick, 2005, Measuring Volatility with the Realized Range, *Journal of Econometrics*, forthcoming.

[43] Meddahi, Nour, 2001, A Theoretical Comparison between Integrated and Realized Volatilities, manuscript, Université de Montréal.

[44] Mincer, Jacob, and Zarnowitz, Victor, 1969, The Evaluation of Economic Forecasts, in Zarnowitz, J. (ed.) *Economic Forecasts and Expectations*, National Bureau of Economic Research, New York.

[45] Newey, Whitney K., and West, Kenneth D., 1987, A Simple, Positive Semidefinite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix, *Econometrica*, 55(3), 703-708.

[46] Pagan, Adrian R., and Schwert, G. William, 1990, Alternative Models for Conditional Volatility, *Journal of Econometrics*, 45, 267-290.

[47] Parkinson, Michael, 1980, The Extreme Value Method for Estimating the Variance of the Rate of Return, *Journal of Business*, 53(1), 61-65.

[48] Patton, Andrew J., and Timmermann, Allan, 2004, Properties of Optimal Forecasts under Asymmetric Loss and Nonlinearity, Centre for Economic Policy Research Discussion Paper 4037.

[49] Poon, Ser-Huang and Granger, Clive W. J., 2003, Forecasting Volatility in Financial Markets, *Journal of Economic Literature*, 41, 478-539.

[50] Shephard, Neil, 2005, *Stochastic Volatility: Selected Readings*, Oxford University Press, United Kingdom.

[51] West, Kenneth D., 2005, Forecast Evaluation, in the *Handbook of Economic Forecasting*, G. Elliott, C.W.J. Granger and A. Timmermann ed.s, North Holland Press, Amsterdam.

[52] West, Kenneth D., Edison, Hali J., and Cho, Dongchul, 1993, A Utility-Based Comparison of Some Models of Exchange Rate Volatility, *Journal of International Economics*, 35, 23-45.

[53] White, Halbert, 1994, *Estimation, Inference and Specification Analysis*, Econometric Society Monographs No. 22, Cambridge University Press, Cambridge, U.K.

[54] White, Halbert, 1999, *Asymptotic Theory for Econometricians,* Revised Edition, Academic Press, San Diego.

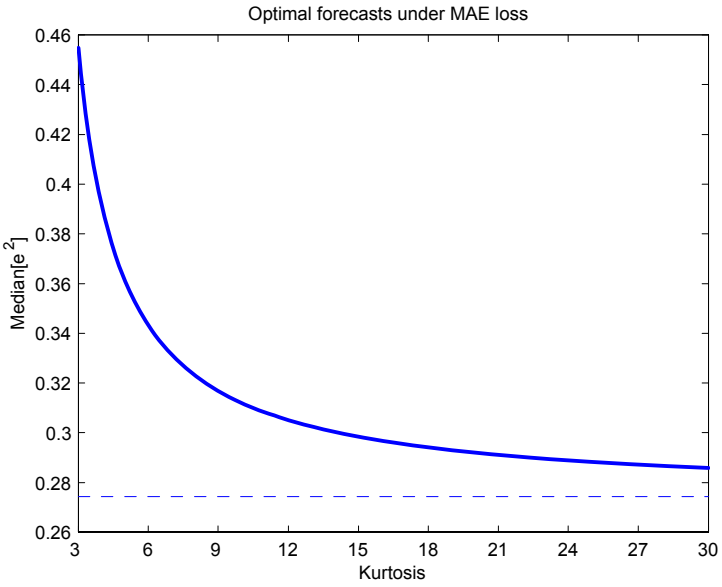

Figure 1: *Optimal forecasts under MAE loss when true variance is 1, for various levels of kurtosis, using the standardised Student's t distribution. The dashed line represents the optimal forecast as* $\nu \to 4$.

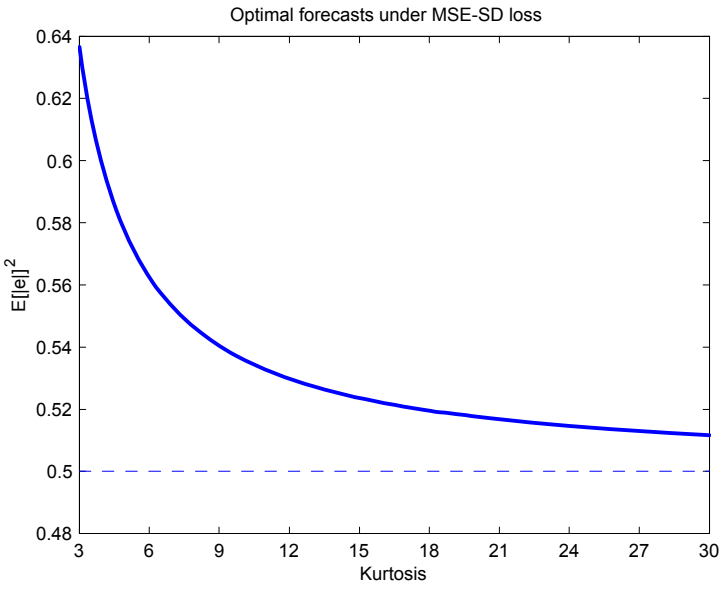

Figure 2: *Optimal forecasts under MSE-SD loss when true variance is 1, for various levels of kurtosis, using the standardised Student's t distribution. The dashed line represents the optimal forecast as* $\nu \to 4$.

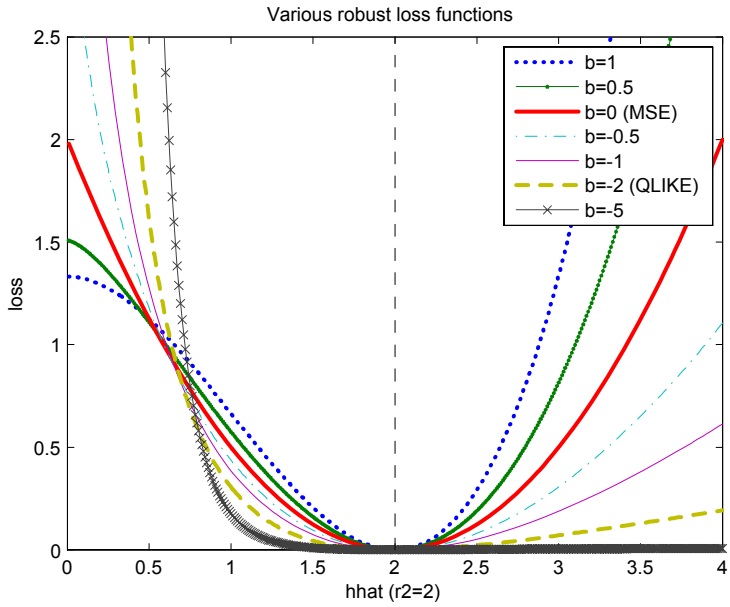

Figure 3: *Loss functions for various choices of b. True* $\hat{\sigma}^2$*=2 in this example, with the volatility forecast ranging between 0 and 4. b=0 and b=-2 correspond to the MSE and QLIKE loss functions respectively.*

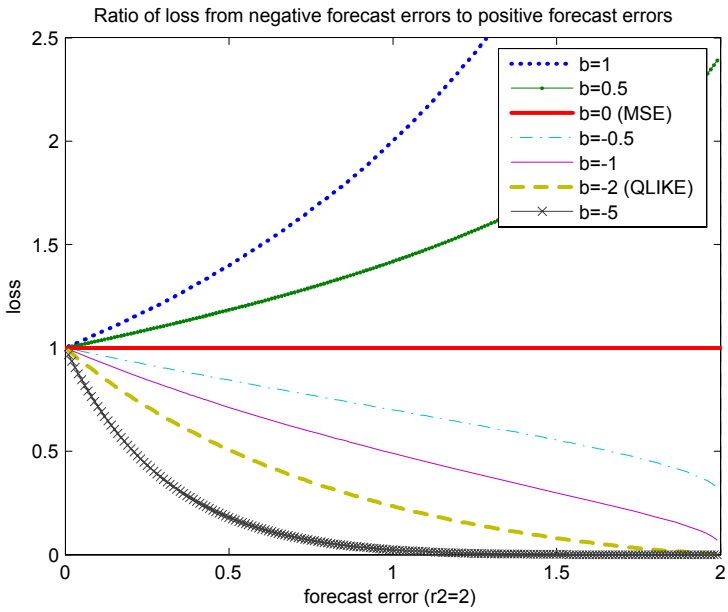

Figure 4: *Ratio of losses from negative forecast errors to positive forecast errors, for various choices of b. True* $\hat{\sigma}^2$*=2 in this example, with the volatility forecast ranging between 0 and 4. b=0 and b=-2 correspond to the MSE and QLIKE loss functions respectively.*

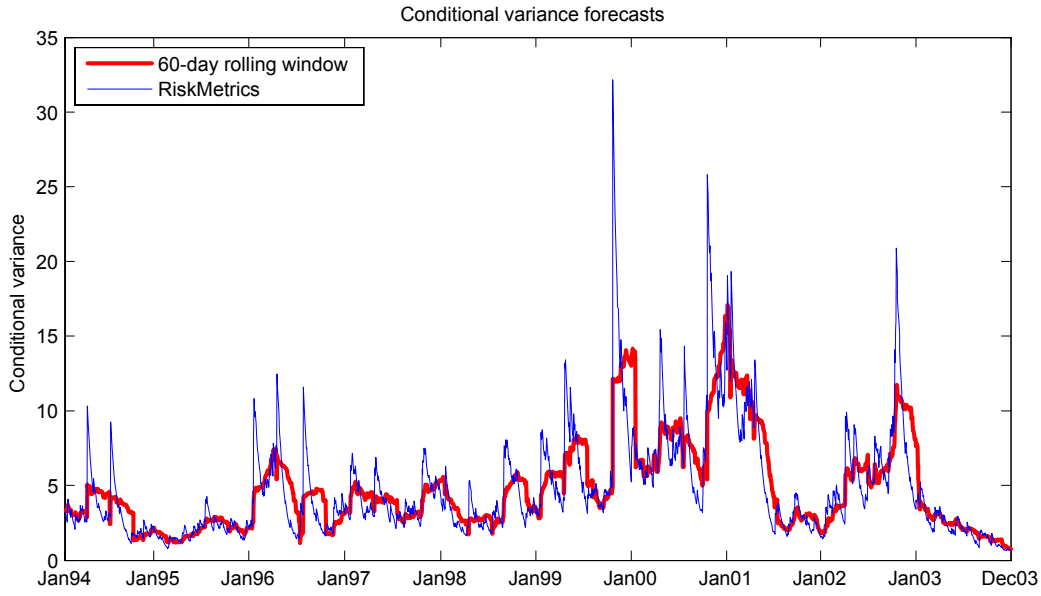

Figure 5: *Conditional variance forecasts for IBM returns from 60-day rolling window and RiskMetrics models, January 1994 to December 2003.*